# Vers une adaptation autonome des modèles acoustiques multilingues pour le traitement automatique de la parole

Sethserey Sam

UNIVERSITÉ DE GRENOBLE

# THÈSE

Pour obtenir le grade de

## DOCTEUR EN SCIENCES DÉLIVRÉ PAR L'UNIVERSITÉ DE GRENOBLE

Spécialité : **MSTII/INFORMATIQUE**

Arrêté ministériel : 7 août 2006


Présentée par

**Sethserey SAM**

Thèse dirigée par **Laurent BESACIER** et
codirigée par **Eric CASTELLI**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
et du **Centre de Recherche International Multimédia, Information, Communication et Applications**
dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique.**


# Vers une adaptation autonome des modèles acoustiques multilingues pour le traitement automatique de la parole

Thèse soutenue publiquement le **07 juin 2011**,
devant le jury composé de :

**M. Hervé Glotin**
Professeur à l'USTV (Toulon), Rapporteur
**M. Christophe Cerisara**
Chercheur/HDR à LORIA (Vandoeuvre-lès-Nancy), Rapporteur
**M. Christian Boitet**
Professeur à l'UJF (Grenoble), Examinateur
**Mme. Martine Adda-Decker**
Directrice de recherche à LIMSI-CNRS (Paris), Examinateur
**M. Laurent Besacier**
Professeur à l'UJF (Grenoble), Membre
**M. Eric Castelli**
Chercheur/HDR à MICA-CNRS (Hanoi), Membre

*គោរពជូនលោកឪពុក និងអ្នកម្តាយ!*

*À mes parents*

*To Mum and Dad*

# Remerciements

Je tiens à remercier Laurent Besacier et Eric Castelli pour m'avoir accueilli, respectivement, au sein du laboratoire LIG (Grenoble, France) et au centre de recherche international MICA (Hanoi, Vietnam) et pour avoir accepté d'être mes deux directeurs de thèse. Un grand remerciement très chaleureusement à Laurent Besacier, qui m'a guidé tout au long de ces années de thèse, pour ses critiques, ses conseils très précis sur mes travaux de recherche et pour avoir relu, corrigé et commenté très soigneusement ce manuscrit. Un grand remerciement également à Eric Castelli pour son aide dévouée sur mes travaux de thèse, surtout la constitution du corpus multilingue, et aussi pour ses conseils très utiles et sa relecture de tout mon manuscrit.

J'adresse mes remerciements à Hervé Glotin et Christophe Cerisara pour avoir accepté d'être rapporteurs de ma thèse. Je voudrais remercier aussi à Christian Boitet pour avoir accepté d'être le président du jury. Je remercie Martine Adda-Decker pour sa participation au jury de cette thèse.

J'adresse mes remerciements à Haizhou Li pour m'avoir accueilli dans son département HLT (*Human Language Technology*) de l'I2R (*Institute Infocom Research, Singapore*) et pour l'intérêt porté à mes travaux de recherche. Un grand remerciement à Cheung Chi Leung et Xiong Xiao, les chercheurs de l'I2R pour leurs aides dévouées, particulièrement sur les travaux présentés dans les deux derniers chapitres de la thèse.

Je tiens à remercier Pham Thi Ngoc Yen, la directrice du Centre MICA (Hanoi, Vietnam) pour m'avoir accueilli pendant les alternances de cette thèse en cotutelle à MICA.

Je tiens à remercier également tous les membres de l'équipe GETALP du laboratoire LIG et du groupe parole du centre de recherche MICA pour leur accueil et leur sympathie. Un grand remerciement à mes amis cambodgiens à Grenoble (Sopheap Seng, Sereysethy Touch, Long Bun, Sophal Ly) avec qui j'ai partagé des grands moments au cours de ma thèse.

Enfin, je voudrais exprimer mes plus profonds remerciements à mes parents, à ma femme, à ma petite sœur, à toute ma grande famille et à Kulthida Sam, ma fille, pour leurs sentiments, leurs soutiens et leurs encouragements pendant tout le temps où j'ai effectué cette thèse.

Un grand merci à tous !

Sethserey SAM

## Résumé

Les technologies de reconnaissance automatique de la parole sont désormais intégrées dans de nombreux systèmes. La performance des systèmes de reconnaissance vocale pour les locuteurs non natifs continue cependant à souffrir de taux d'erreur élevés, en raison de la différence entre la parole non native et les modèles entraînés. La réalisation d'enregistrements en grande quantité de parole non native est généralement une tâche très difficile et peu réaliste pour représenter toutes les origines des locuteurs.

Cette thèse porte sur l'amélioration des modèles acoustiques multilingues pour la transcription phonétique de la parole de type « réunion multilingue ». Traiter ce type de parole implique de relever plusieurs défis : 1) il peut exister de la conversation entre des locuteurs natifs et non natifs ; 2) il y a non seulement de la parole non native d'une langue, mais de plusieurs langues parlées par des locuteurs venant de différentes origines ; 3) il est difficile de collecter suffisamment de données pour amorcer les systèmes de transcription. Pour résoudre ces problèmes, nous proposons un processus d'adaptation de modèles acoustiques multilingues que nous appelons « adaptation autonome ». Dans l'adaptation autonome, nous étudions plusieurs approches pour adapter les modèles acoustiques multilingues de manière non supervisée (les langues parlées et les origines des locuteurs ne sont pas connues à l'avance) en n'utilisant aucune donnée supplémentaire lors du processus d'adaptation. Les approches étudiées sont décomposées selon deux modules. Le premier module s'appelle « l'observateur de langues » et a pour but de calculer les caractéristiques linguistiques (langues parlées et origines des locuteurs) des segments à décoder. Le deuxième module vise à adapter le modèle acoustique multilingue en fonction des connaissances fournies par l'observateur de langues. Pour évaluer l'utilité de l'adaptation autonome d'un modèle acoustique multilingue, nous utilisons des données de test, qui sont extraites de réunions multilingues, contenant de la parole native et non native de trois langues : l'anglais (EN), le français (FR) et le vietnamien (VN). Selon les résultats d'expérimentation, l'adaptation autonome donne des résultats prometteurs pour la parole non native, mais dégrade très légèrement les performances sur de la parole native. Afin d'améliorer la performance globale des systèmes de transcription pour la parole native ainsi que non native, nous étudions plusieurs approches de détection de la parole non native et proposons de mettre un tel détecteur en cascade avec notre processus d'adaptation autonome. Les résultats obtenus ainsi sont les meilleurs parmi toutes les expériences réalisées sur notre corpus de réunions multilingues.



**Mots-clés :** Reconnaissance de la parole non native, adaptation autonome de modèles acoustiques multilingues, observateur de langues, interpolation, discrimination entre parole native et non native.



## Abstract

Automatic speech recognition technologies are now integrated into many systems. The performance of speech recognition systems for non-native speakers, however, continues to suffer from high error rates, due to the difference between non-native speech and models trained on native speech. The making of recordings in large quantities of non-native speech to represent all the origins of the speakers is a very difficult and impractical task.

This thesis focuses on improving multilingual acoustic models for automatic phonetic transcription of speech in "multilingual meetings". There are several challenges in "multilingual meeting" speech: 1) there can be a conversation between native and non-native speakers; 2) there is not only one language spoken by a non-native, but several languages spoken by speakers from different origins; 3) it is difficult to collect sufficient data to bootstrap the transcription systems. To meet these challenges, we propose a process of adaptation of multilingual acoustic models called "autonomous adaptation". In autonomous adaptation, we studied several approaches for adapting multilingual acoustic models in an unsupervised way (spoken languages and speakers' origins are not known in advance) and no additional data are used during the adaptation process. The approaches studied are decomposed into two modules. The first module, called the "language observer", recovers the linguistic information (spoken languages and speakers' origins) of the segments to be decoded. The second module adapts the multilingual acoustic models based on knowledge provided by the language observer. To evaluate the usefulness of autonomous adaptation of multilingual acoustic models, we use a set of test data, which are extracted from multilingual meeting corpora, containing native and non-native speech in three languages: English (EN), French (FR) and Vietnamese (VN). According to the experimental results, the autonomous adaptation approach shows promising results for non-native speech, but degrades very slightly the performance on native speech. To improve the overall performance of transcription systems for both native and non-native speech, we study several approaches for detecting non-native speech, and propose to cascade such a detector with our self-adaptation process (autonomous adaptation). The results obtained so far are the best among all experiments done on our corpus of multilingual meetings.



**Keywords:** Non-native speech recognition, autonomous adaptation of multilingual acoustic models, language observer, interpolation, discrimination between native and non-native speech.

# Table des matières

# Liste des figures

# Liste des tableaux

# Introduction

De nos jours où la concurrence des marchés devient de plus en plus mondiale et critique, les personnes qui montrent des capacités à communiquer en plusieurs langues ont plus de facilités pour développer leurs projets personnels ou professionnels. En effet, la langue ne joue pas seulement un rôle fondamental dans la communication entre les individus, mais elle représente aussi l'identité et la culture des communautés à laquelle ils appartiennent.

Par ailleurs, les *lingua franca* (les langues largement utilisées) comme l'anglais, l'espagnol et le français, sont depuis longtemps privilégiées dans le monde entier en raison de leur richesse, de leur rôle historique, mais aussi de leur utilisation forte et constante dans le domaine de la science et de la technologie. Par conséquent, ces langues sont enseignées dans la plupart des écoles et universités du monde entier.

De plus, l'homme se déplace et migre beaucoup plus qu'auparavant. Par exemple, aux États-Unis, le pourcentage de personnes d'origine étrangère a été multiplié par trois en huit ans entre 1998 et 2006 (7.3 % en 1998 et 24.2 % en 2006) [Ohlemacher, 2007]. Par ailleurs, le tourisme est une industrie très lucrative pour de nombreux pays. En France, par exemple, il y a eu 78 millions de touristes qui ont visité le pays en 2006 [L'Expansion, 2007].

Aujourd'hui, avec le développement très rapide des technologies vocales, l'utilisation de systèmes de reconnaissance automatique de la parole (RAP) en tant qu'interfaces naturelles est de plus en plus courante. Beaucoup d'applications de RAP sont intégrées dans différents types de systèmes, tels que les ordinateurs, les téléphones portables, les voitures, les machines, etc.

Cependant, les utilisateurs non natifs d'applications de RAP sont souvent découragés par les performances des systèmes. Des études [Liu, 2006 ; Oh, 2007 ; Tan, 2008 ; Wang, 2003] montrent que la performance des systèmes de reconnaissance vocale sur de la parole non native est au moins deux fois inférieure à celle obtenue dans le cas de la parole native.

Les systèmes statistiques de reconnaissance automatique de la parole (RAP) utilisent principalement trois types de modélisation, à savoir la modélisation acoustique, la modélisation lexicale et la modélisation de la langue. Ces modèles sont créés en utilisant une approche dite empirique (*data-driven*). Comme ils sont généralement générés en utilisant seulement des ressources (texte et signaux) de parole native, il peut y avoir une forte disparité

entre la parole non native des locuteurs utilisateurs et celle (native) utilisée pour réaliser ces trois modèles. En conséquence, le taux de reconnaissance de la parole non native est beaucoup plus faible que celui de la parole native.

Une meilleure solution consisterait à construire des modèles qui correspondent mieux à des locuteurs non natifs en utilisant énormément de ressources (textes et signaux) de parole non native. Toutefois, l'acquisition de telles ressources prend beaucoup de temps et demande d'importants moyens financiers, car une langue peut être parlée par des locuteurs non natifs venant de multiples pays: des corpus intégrant des locuteurs de chaque pays sont nécessaires. Dans certain cas, il est même très difficile, voire impossible, de réaliser ces corpus, comme par exemple dans le cas des langues peu dotées. Cet état de fait constitue un verrou important quant au déploiement des systèmes de reconnaissance automatique de la parole non native.

Cette thèse porte sur l'amélioration des modèles acoustiques multilingues pour la reconnaissance automatique de la parole de type « réunion multilingue ». Traiter ce type de parole implique de relever plusieurs défis : 1) dans les réunions multilingues, il peut y avoir de la conversation entre les locuteurs natifs et non natifs ; 2) il y a non seulement de la parole non native d'une langue, mais aussi plusieurs langues parlées par des locuteurs venant de différentes origines ; 3) un locuteur peut mélanger plusieurs langues dans son discours (*code-mixing*). En outre, il est difficile (voir impossible) de collecter suffisamment de données pour amorcer les systèmes de RAP pour tout type de parole pouvant être rencontré lors de « réunions multilingues ».

Des études [Flege, 1997 ; Huang, 2001] sur la parole non native montrent, ce qui semble évident, que les locuteurs empruntent des caractéristiques acoustiques de leur langue maternelle quand ils parlent d'autres langues. D'ailleurs, pour améliorer la performance d'un système de RAP sur de la parole non native, une technique d'interpolation entre deux modèles acoustiques (celui de la langue maternelle du locuteur ou « langue source » et celui de la langue parlée ou « langue cible ») a été proposée dans une thèse précédente au LIG [Tan, 2007]. Cependant, la solution proposée est limitée au cas où la langue parlée et l'origine du locuteur sont connues à l'avance.

Suite à ces dernières études, et pour répondre aux défis que représente l'étude de la parole dans le cas de réunions multilingues, nous proposons dans cette thèse une adaptation autonome du modèle acoustique multilingue pour améliorer la performance du système de transcription phonétique de la parole utilisé pour les réunions multilingues. L'adjectif « autonome » signifie que le modèle acoustique (multilingue) est automatiquement réadapté

pour mieux décoder le flux des signaux inconnus (aucune information sur la langue parlée et l'origine du locuteur n'est supposée connue à l'avance) sans besoin de données adaptées supplémentaires.

Dans le chapitre 1 de ce manuscrit, après une brève présentation de la motivation et de la difficulté rencontrée pour traiter la parole non native pour des locuteurs de différentes origines, nous présentons l'architecture de l'adaptation autonome proposée. Puis, nous présentons le principe général de la reconnaissance automatique de la parole par modèles statistiques et dégageons les avantages des modèles acoustiques multilingues. Ces derniers peuvent être utilisés pour traiter des problèmes comme la reconnaissance des langues peu dotées, le *code-mixing*, la recherche d'informations multilingue et aussi la reconnaissance de la parole non native. Nous faisons également référence, dans cette partie, à notre étude récente sur l'utilisation de modèles acoustiques multilingues pour la transcription automatique d'une langue en danger (le mo piu, cette étude est détaillée dans l'annexe C du manuscrit car elle nous a semblé déconnectée du reste de la thèse). Nous présentons par la suite une discussion sur l'acquisition de la langue source et de la langue cible chez l'être humain ; nous expliquons également les principaux facteurs de dégradation de la RAP sur de la parole non native. À la fin du chapitre, des méthodes d'adaptation standard dans le domaine de la reconnaissance de la parole non native sont présentées.

Pour évaluer l'utilité de l'adaptation autonome d'un modèle acoustique multilingue, il est nécessaire d'avoir des données de test contenant au moins de la parole native et non native des langues traitées (dans notre cas, les trois langues anglaise (EN), française (FR) et vietnamienne (VN)). Le chapitre 2 décrit, étape par étape, le processus utilisé pour acquérir notre corpus de type « réunion multilingue », nommé « MICAMultiMeet », à partir duquel les données de test ont été extraites. Ce corpus contient de la parole native de quatre langues (l'anglais, le français, le vietnamien et le khmer) et de la parole non native de trois langues (l'anglais, le français et le vietnamien).

Dans le chapitre 3, nous détaillons deux approches pour construire ce que nous avons appelé un « observateur de langue », le module cœur de l'adaptation autonome. La première approche génère les scores postérieurs des langues à partir d'un décodage acoustico-phonétique. La deuxième approche produit les scores postérieurs des langues selon un modèle phonotactique.

Le chapitre 4 présente les différentes méthodes d'adaptation du modèle acoustique multilingue. Ces méthodes n'utilisent que les signaux à décoder dans leur processus

d'adaptation, et elles sont divisées en deux groupes : 1) l'adaptation « en ligne » est faite en utilisant seulement la phrase en cours de décodage (phrase courante) ; 2) l'adaptation « hors ligne » est faite en utilisant la phrase courante ainsi que les phrases déjà décodées.

Dans le chapitre 5, nous étudions différentes approches de discrimination entre parole native et non native. Dans cette étude, encore préliminaire, nous présentons les résultats de détection sur la parole native et non native française. À la fin du chapitre, nous utilisons ces systèmes de détection au début de la chaîne du processus d'adaptation autonome et nous comparons la performance de ce dernier processus d'adaptation avec le processus d'adaptation autonome seul.

Le document se termine par la conclusion globale de la thèse, suivie de nos perspectives à court et long terme.

# Chapitre 1 :
# Contexte de l'étude et état de l'art

## 1.1. Contexte de l'étude

### 1.1.1. Parole non native dans le monde

Les langues comme l'anglais, le chinois, l'espagnol et le français sont parlées dans le monde entier. Par exemple, d'après la base de donnée "The British Council" [Graddol, 1997], il y aurait, en 2007, plus d'un milliard de locuteurs non natifs anglais répartis sur plus de 180 pays dans le monde, alors que les locuteurs natifs anglais représentent seulement 375 millions de personnes. Dans le tableau 1.1 qui présente des études des langues parlées en Europe [Lazzari, 2006], nous remarquons que le nombre de locuteurs non natifs est aussi important que celui des locuteurs natifs pour presque toutes les langues listées.

| **Langue** | **parlée comme langue maternelle** | **parlée comme langue étrangère** | **pourcentages en Europe** |
|---|---|---|---|
| Anglais | 13 % | 34 % | 47 % |
| Allemand | 18 % | 12 % | 30 % |
| Français | 12 % | 11 % | 23 % |
| Italien | 13 % | 2 % | 15 % |
| Espagnol | 9 % | 5 % | 14 % |
| Polonais | 9 % | 1 % | 10 % |
| Néerlandais | 5 % | 1 % | 6 % |
| Russe | 1 % | 5 % | 6 % |

*Tableau 1.1 : Langues parlées en Europe [Lazzari, 2006]*

Dans ce manuscrit, nous utilisons le terme « langue source » ou L1 pour parler de la langue maternelle ou la langue d'origine des locuteurs et le terme « langue cible » ou L2 pour parler de la langue pour laquelle nous souhaitons développer un système de reconnaissance automatique de la parole.

### 1.1.2. Les besoins des systèmes de traitement automatique de la parole non native

Au cours des dernières années, il y a eu beaucoup de progrès dans le domaine de la reconnaissance automatique de la parole (RAP) et des systèmes de dialogue. Cela rend les systèmes de traitement de la parole prêts pour des applications intégrées sur de multiples terminaux : ordinateurs personnels, téléphones portables, voitures, etc.

Cependant, le problème du traitement de la parole non native reste toujours un défi important pour les systèmes de reconnaissance de la parole. Des études montrent que la performance des systèmes de RAP non native est au moins deux fois inférieure à celle de la parole native. L'une des principales causes de la dégradation des performances de systèmes de reconnaissance est que leurs trois modèles (les modélisations acoustique, de langue et lexicale) sont entraînés sur de la parole native. Concernant les modèles acoustiques, il existe plusieurs solutions : 1) amorcer les modèles des systèmes RAPs en utilisant une grande base de données non native (textes et parole) ; 2) réadapter des modèles natifs en utilisant très peu de données ou des informations linguistiques issus de parole non native. La première est, en théorie, la meilleure solution, mais l'acquisition de ressources de parole non native prend beaucoup de temps et s'avère quasiment impossible si l'on souhaite considérer une langue parlée par des locuteurs non natifs venant du monde entier.

| **Références** | **Langue cible (L2)** | **Langue source (L1)** | **WER (native)** | **WER (non native)** |
|---|---|---|---|---|
| [Tan, 2008] | Français | Vietnamien/Mandarin | 11,9 % | 58,5 % / 60,6 % |
| [Oh, 2007] | Anglais | Coréen | 4,2 % | 39,2 % |
| [Liu, 2006] | Mandarin | Cantonais | 7,9 % | 20,0 % |
| [Wang, 2003] | Anglais | Allemand | 16,2 % | 49,3 % |
| [Witt, 1999] | Anglais | Espagnol/Japonais | - | 28,2 % / 28,2 % |

*Tableau 1.2 : Comparaison des performances de systèmes de RAP sur différents locuteurs non natifs*

Concernant les recherches très récentes sur la RAP non native, nous remarquons également que les études se limitent à 1) des langues cibles bien dotées (telles que l'anglais, le français, etc.) ; 2) une langue cible parlée par des locuteurs non natifs venant d'une ou deux origines (langue source) maximum. Le tableau 1.2 présente des résultats intéressants sur des systèmes RAPs non natives.

### 1.1.3. Travail de thèse

Ce travail de thèse s'inscrit dans un souci constant d'améliorer la performance des systèmes de reconnaissance acoustico-phonétique de parole de type « réunion multilingue ». Pour traiter ce type de parole, plusieurs problèmes se présentent :

- il existe plusieurs langues cibles parlées par des locuteurs de différentes origines ;
- il existe de la parole native et non native dans la conversation ;
- les locuteurs peuvent alterner plusieurs langues dans un même discours (code-switching).

Dans notre contexte d'étude, nous considérons principalement le premier et le deuxième problème. Dans l'approche proposée, nous nous sommes donné comme contrainte de n'utiliser aucune donnée supplémentaire lors du processus d'adaptation, sauf le modèle multilingue contenant au moins les modèles acoustiques des langues sources et ceux des langues cibles visées.

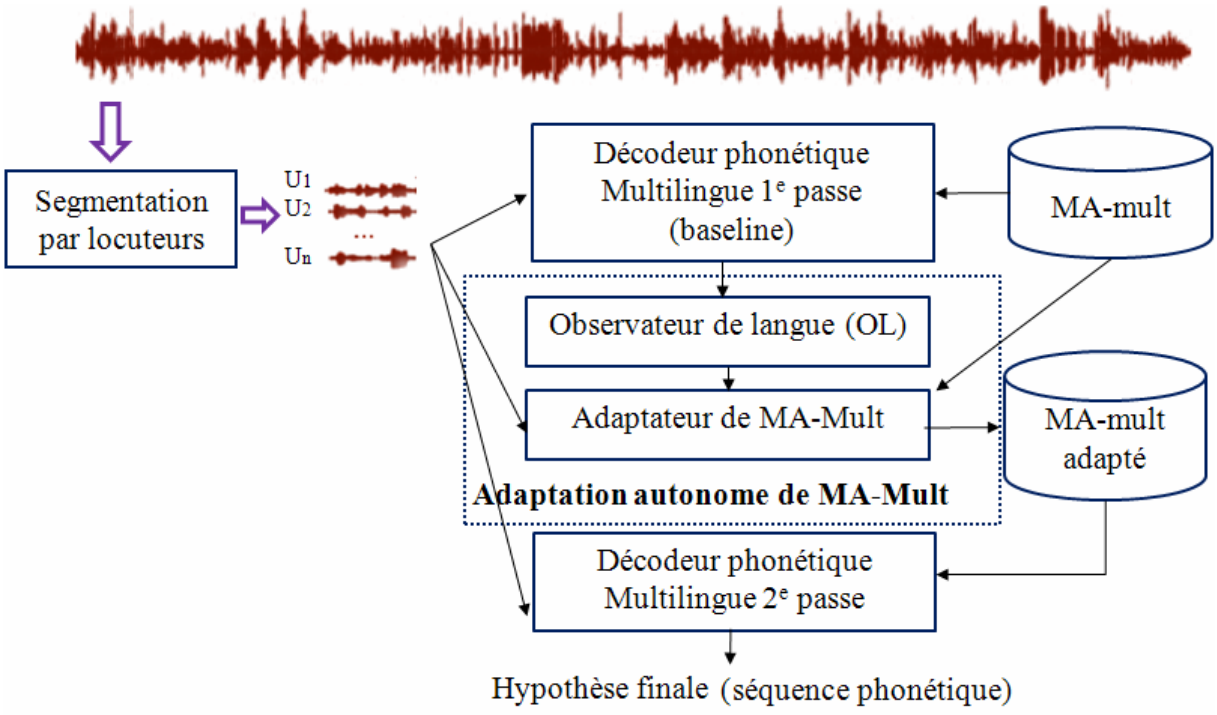



*Figure 1.1 : Processus d'adaptation autonome du modèle acoustique multilingue*

Nous proposons une méthode d'adaptation non supervisée d'un modèle acoustique multilingue pour le décodage acoustico-phonétique multilingue. Cette approche est qualifiée d'*autonome* car le modèle acoustique (multilingue) est automatiquement réadapté au cours du décodage pour mieux transcrire le flux de signaux inconnus (aucune information sur la langue parlée et l'origine du locuteur n'est supposée connue à l'avance) sans besoin de données

adaptées supplémentaires. Pour atteindre cet objectif, nous proposons une architecture en deux passes comme illustré dans la figure 1.1 [Sam, 2010a].

Dans la figure 1.1, le flux des signaux est, tout d'abord, segmenté (manuellement ou automatiquement) selon les changements de locuteur. Ensuite, les segments de parole sont décodés par le décodeur acoustico-phonétique qui n'utilise que le modèle acoustique multilingue (notre système *baseline*). Puis, le module « observateur de langue » (OL) capture les informations sur la langue parlée et l'origine du locuteur à partir de l'hypothèse du système *baseline*. Ensuite, l'interpolation entre plusieurs modèles acoustiques est faite en utilisant les scores postérieurs de langue générés par l'observateur (OL). Enfin, nous utilisons le modèle acoustique multilingue interpolé lors du décodage acoustico-phonétique multilingue de la deuxième passe.

Les motivations pour ce type d'adaptation de modèle acoustique sont que : 1) nous pourrons connaître la langue parlée et, peut être, aussi l'origine du locuteur en utilisant le module « observateur de langue » ; 2) en plus, avec l'approche d'interpolation de plusieurs langues, nous n'avons besoin d'aucune donnée supplémentaire pour réadapter le modèle acoustique multilingue.

En outre, nous appliquons les approches proposées à la parole native et non native de trois langues cible et source (l'anglais, le français et le vietnamien) sachant que l'une des trois langues est une langue peu dotée (le vietnamien). Il est important de mentionner qu'une langue peu dotée [Berment, 2004] est définie comme une langue qui ne possède pas encore ou pas beaucoup (en quantité et en qualité) de ressources linguistiques pour la construction des systèmes de traitement de langue, par exemple des systèmes de reconnaissance automatique de la parole, ce qui est un problème particulier dans un contexte d'apprentissage statistique où les données doivent être disponibles en grande quantité.

Dans les sections suivantes, nous présentons le principe de la reconnaissance de la parole par l'approche statistique, le principe des modèles acoustiques multilingues, et les études récentes sur la reconnaissance automatique de la parole non native. Le corpus multilingue utilisé dans nos études, ainsi que les modules de l'adaptation autonomes, seront détaillés dans les chapitres suivants.

# 1.2. Reconnaissance automatique de la parole monolingue

## 1.2.1. Historique

La reconnaissance automatique de la parole (RAP) consiste à extraire, à l'aide d'un ordinateur, l'information lexicale contenue dans un signal de parole.

Les fondements de la technologie récente en reconnaissance de la parole ont été élaborés par F. Jelinek et son équipe à IBM dans les années 70 [Jelinek, 1976]. Les premiers travaux (années 80) se sont intéressés aux mots, et ce, pour des applications à vocabulaire réduit. Au début des années 90, les systèmes de reconnaissance automatique de parole continue à grand vocabulaire et indépendants du locuteur ont vu le jour. La technologie s'est développée rapidement et, déjà vers le milieu des années 90, une précision raisonnable était atteinte pour une tâche de dictée vocale. Une partie de ce développement a été réalisée dans le cadre de programmes d'évaluation de la DARPA (*Defense Advanced Research Projects Agency*).

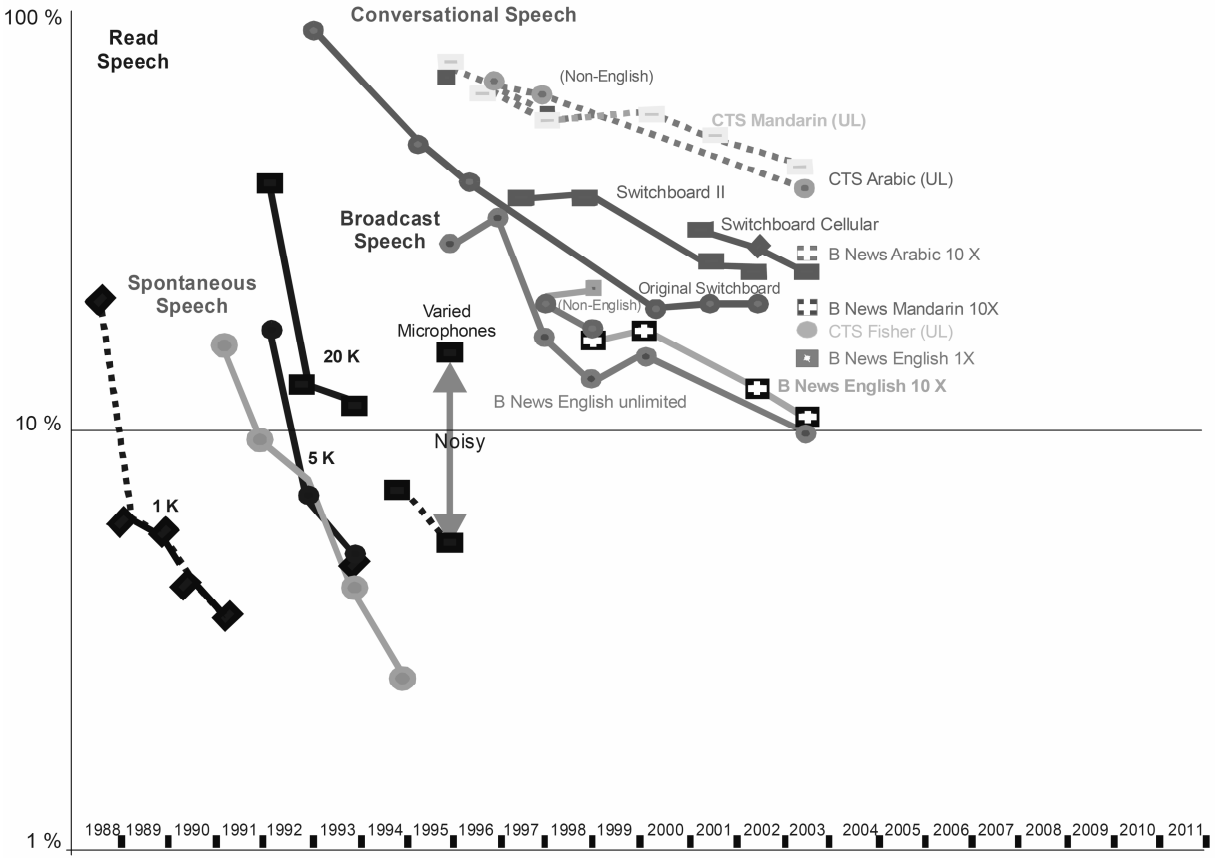



*Figure 1.2 : Evolution des performances (taux d'erreurs) de RAP (NIST [Pallett, 2003])*

Différents systèmes de reconnaissance de la parole ont été développés, couvrant des domaines variés : reconnaissance de différents types de parole (téléphonique, continue, mots isolés, etc.), systèmes de dictée vocale, systèmes de commande et contrôle sur PC, systèmes

de compréhension en langage naturel. La figure 1.2 représente 15 ans d'amélioration des performances pour différentes tâches de reconnaissance de la parole, dans le contexte des campagnes d'évaluation organisées par l'organisme américain NIST [Pallett, 2003].

### 1.2.2. Formulation du processus de décodage en RAP

Les premiers travaux de reconnaissance de la parole ont essayé d'appliquer des connaissances expertes en production et en perception. De nos jours, les techniques de modélisation statistique apportent encore les meilleures performances.

En général, les systèmes de reconnaissance automatique de la parole, à base de modélisation statistique, suivent le schéma représenté par la figure 1.3.

À partir d'un signal de parole, le premier traitement consiste à extraire les paramètres caractéristiques. Ces paramètres sont mis en entrée d'un modèle acoustique, ou de décodage acoustico-phonétique. Ce décodage acoustico-phonétique peut produire à son tour une ou plusieurs hypothèses phonétiques associées en général à une probabilité pour chaque segment (une fenêtre ou une trame) de signaux de la parole. Ce générateur d'hypothèses locales est souvent modélisé par des modèles statistiques d'unités élémentaires de parole, par exemple un phonème. Pour entraîner des modèles acoustiques, nous apprenons les modèles des unités acoustiques sur un grand corpus étiqueté.

Le générateur d'hypothèses interagit avec un modèle lexical pour forcer le décodage acoustico-phonétique à ne reconnaître que des mots représentés dans le modèle lexical. Le modèle lexical est représenté par un dictionnaire de prononciation (dictionnaire phonétique) ou par des automates probabilistes qui sont capables d'associer une probabilité à chaque prononciation possible d'un mot.

Pour la reconnaissance automatique de la parole continue à grand vocabulaire, le générateur interagit avec un module syntaxique (le plus souvent un modèle de langage probabiliste qui n'a de syntaxique que le nom) pour forcer le reconnaisseur à intégrer des contraintes sur l'enchaînement entre les mots. Ces contraintes sont souvent formalisées par des modèles de langage. Pour reconnaître ce qui est dit, on commence par chercher, grâce aux modèles d'unités acoustiques, l'unité qui est supposée avoir été produite, puis on construit, à partir du treillis d'unités acoustiques et d'un modèle statistique du langage, la suite de mots la plus probable.

Nous donnons l'équation bayésienne appliquée au problème du décodage de reconnaissance automatique de la parole. Soient $o$ une séquence de vecteurs acoustiques

inconnus et $m_i$ ($i=1..K$) une parmi $K$ classes possibles pour cette observation (ex. phonèmes, mots, etc.). La classe reconnue est :

$$m^* = \text{argmax}\ \frac{P(o/m_i).P(m_i)}{P(o)} = \text{argmax}\ P(o/m_i).P(m_i) \qquad (1.1)$$

Le mot reconnu $m^*$ sera donc celui qui maximise cette quantité, parmi tous les mots candidats $m_i$. Les probabilités $P(o|m_i)$ d'observer le signal $o$ connaissant la séquence $m_i$ nécessitent un modèle acoustique pour être estimées. Les probabilités $P(m_i)$ a priori de la séquence, indépendamment du signal, nécessitent un modèle de langage pour être estimées. $P(o)$ est la probabilité du signal qui est la même pour toutes les séquences possibles donc sa valeur n'est pas prise en compte.

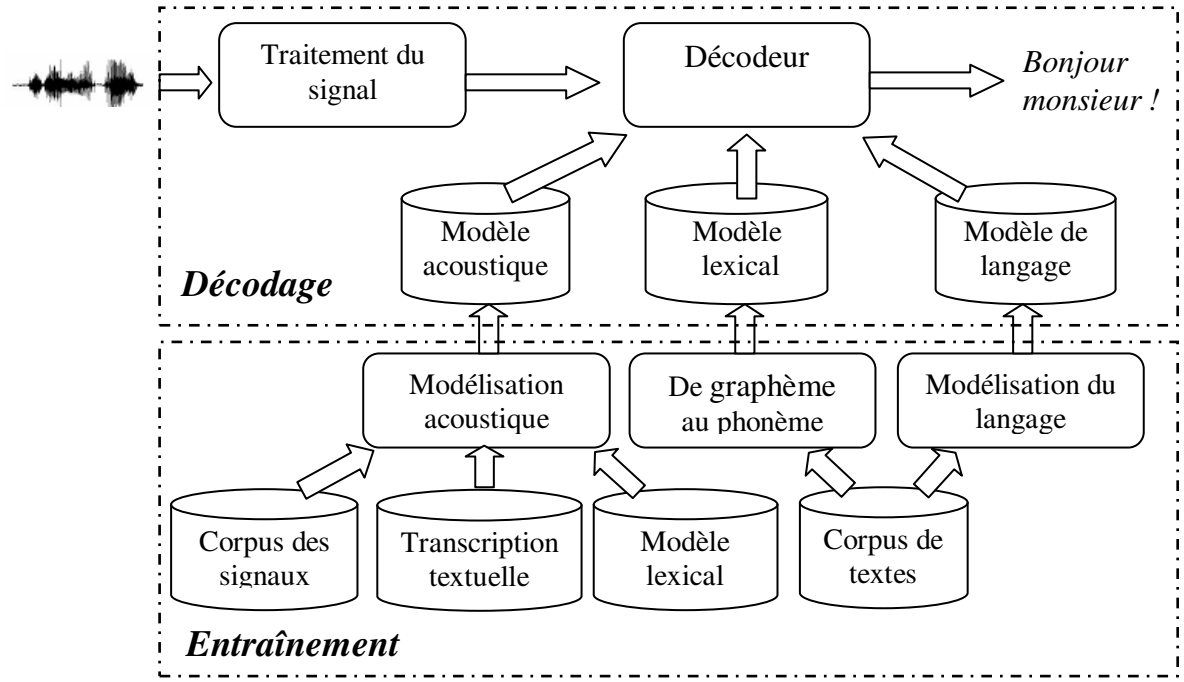



*Figure 1.3 : Architecture globale d'un système reconnaissance automatique de la parole (RAP) par modélisation statistique*

Avant de présenter les principes généraux des différents modules constituant un système de reconnaissance automatique de la parole − modélisation acoustique (décodage acoustico-phonétique), dictionnaires phonétiques, modélisation statistique du langage − nous présentons d'abord les différentes techniques d'évaluation d'un système de reconnaissance automatique de la parole.

### 1.2.3. Évaluation

Les systèmes de reconnaissance de la parole sont évalués en terme de taux de mots erronés (*WER : Word Error Rate*). Généralement, il y a trois types d'erreurs sur les mots reconnus par le système de reconnaissance de la parole :

- substitution (*Sub*) ou remplacement du mot correct par un autre mot ;
- suppression (*Sup*) ou omission d'un mot correct ;
- insertion (*Ins*) ou ajout d'un mot supplémentaire.

Ces trois types d'erreur peuvent être calculés après alignement dynamique entre l'hypothèse du décodeur et une transcription de référence, à l'aide d'un calcul de distance d'édition minimale entre mots. Le résultat sera le nombre minimal d'insertions, de substitutions et d'élisions de mots, pour pouvoir faire correspondre les séquences de mots de l'hypothèse et de la référence. D'après sa définition, le WER peut être supérieur à 100 % à cause des insertions.

$$WER = \frac{nb\ de\ sub + nb\ de\ sup + nb\ de\ ins}{nb\ de\ mots\ corrects\ dans\ la\ référence} \tag{1.2}$$

### 1.2.4. Décodage acoustico-phonétique

Un décodage acoustico-phonétique (DAP) est défini, d'après [Haton, 1991], comme la transformation de l'onde vocale en unités phonétiques ; c'est une sorte de transcodage qui fait passer d'un code acoustique à un code phonétique ou plus exactement comme la mise en correspondance du signal et d'unités phonétiques prédéfinies pour lequel le niveau de représentation passe de continu à discret.

Le décodage acoustico-phonétique est composé d'une première partie consistant à extraire les paramètres acoustiques et à les représenter sous forme de vecteurs acoustiques à partir du signal à décoder, et d'une seconde partie qui, à partir de ces jeux de paramètres, apprend des modèles d'unités acoustiques ou décode le signal d'entrée, selon que l'on veuille apprendre ou reconnaître.

#### *1.2.4.1. Du signal aux vecteurs acoustiques*

Le signal de parole ne peut être exploité directement. En effet, il contient non seulement le message linguistique, mais aussi de nombreux autres éléments comme des informations liées au locuteur, aux conditions d'enregistrement, etc. Toutes ces informations ne sont pas nécessaires lors du décodage de parole et rajoutent même du bruit. De plus, la variabilité et la

redondance du signal de parole le rendent difficilement exploitable tel quel. Il est donc nécessaire d'en extraire uniquement les paramètres qui seront dépendants du message linguistique.

Généralement, ces paramètres sont estimés via des fenêtres glissantes sur le signal. Cette analyse par fenêtrage permet d'estimer le signal sur une portion jugée quasi-stationnaire, généralement 10 à 20 ms, en limitant les effets de bord et les discontinuités du signal via une fenêtre de *Hamming*. La majorité des paramètres représente le spectre fréquentiel et son évolution sur une fenêtre de taille donnée. Les techniques de paramétrage les plus utilisées sont PLP (*Perceptual Linear Prediction dans le domaine spectral*) [Hermansky, 1991], LPCC (*Linear Prediction Cepstral Coefficients dans le domaine temporel*) [Markel, 1982] et MFCC (*Mel Frequency Cepstral Coefficients dans le domaine cepstral*). La paramétrisation n'étant pas au cœur de notre travail de thèse, nous ne détaillons pas plus cette partie.

### *1.2.4.2. Modélisation acoustique à base de modèles de Markov cachés*

Pour la modélisation statistique acoustique, les modèles de Markov cachés (*Hidden Markov model*, HMM) sont aujourd'hui utilisés dans un très grand nombre des systèmes de reconnaissance automatique de la parole. Ces modèles furent introduits par [Baker, 1975; Jelinek, 1976]. Chaque unité acoustique, en effet, est modélisée par un HMM. Dans le cas de petits lexiques, ces unités peuvent être les mots. Dans le cas de grands lexiques, l'unité préférée est le phonème (ou polyphone) ce qui limite le nombre de paramètres à estimer. Dans ce dernier cas, lors de la reconnaissance, les mots sont construits (dynamiquement) en termes de séquences de phonèmes et les phrases en termes de séquences de mots.

Dans le cadre des systèmes de RAP markoviens, chaque unité acoustique est modélisée par un modèle de Markov caché (HMM) typé gauche-droit avec trois à cinq états, comme illustré par la figure 1.4, dans lequel on ne peut pas revenir à un état précédent.

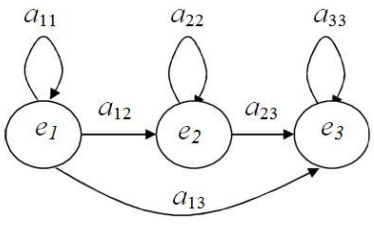



*Figure 1.4 : Exemple de HMM à trois états gauche-droit*

A chaque état du modèle de Markov est associée une distribution de probabilités modélisant la génération des vecteurs acoustiques à partir de cet état. Un HMM est caractérisé par plusieurs paramètres :

- son nombre d'états $N$ ;
- l'ensemble des états du modèle $e = (e_i)_{(1 \le i \le N)}$ ;
- un matrice de transition entre les états $A = (a_{ij})_{(1 \le i,j \le N)}$ avec $a_{ij} = P(e_t = e_j \mid e_{t-1} = e_i)$ ;
- la probabilité d'occupation d'un état à l'instant initial $(\pi_i)_{1 \le i \le N} : \pi_i = P(e_1 = e_i)$ ;
- la densité de probabilité d'observation $o_k$ associée à l'état $e_i$ ; elle dépend du type de HMM :
  - pour un HMM discret, l'émission de probabilités $b_i(k)$ est définie par $b_i(k) = P(x_t = o_k \mid e_t = e_i)_{(1 \le i \le N)}$ ;
  - pour un HMM continu, l'émission de probabilités dans un espace continu $b_i(x)$ est généralement définie par des fonctions de densité multi-gaussiennes (*Gaussian mixture*) : $b_i(x) = \sum_{g=1}^{G_i} c_{ig}.N(x \mid \mu_{ig}, \Sigma_{ig})$, $\sum_{g=1}^{G_i} c_{ig} = 1$ ; où $G_i$ désigne le nombre de composantes des distributions gaussiennes (*Gaussian mixture*) et $c_{ig}$ est le poids (*mixture weights or distribution weights*) associé à la $g^{\text{ème}}$ distribution de l'état $e_i$. Les composants gaussiens (*codebook weights*) $N(x \mid \mu, \Sigma)$ sont généralement définis par un vecteur moyen (*mean vector*) $\mu$ et une matrice de covariance $\Sigma$.

Un HMM est donc représenté par un ensemble de paramètres :

$$HMM = (N, A, \{\pi\}, \{b\}) \tag{1.3}$$

Soit un modèle de Markov caché, il y a trois problèmes fondamentaux à résoudre dans un système de reconnaissance acoustico-phonétique par modélisation statistique :

- Évaluation : soient un modèle $\Phi$ et une séquence d'observations $X = \{x_1, x_2, \ldots, x_T\}$.

  ➔ Comment calculer $P(X|\Phi)$, la probabilité que la séquence des observations ait été émise par le modèle $\Phi$ ?

- Décodage : soient un modèle $\Phi$ et une séquence d'observations $X = \{x_1, x_2, \ldots, x_T\}$.

  ➔ Comment déterminer la séquence d'états cachés $Q = \{q_0, q_1, \ldots, q_T\}$ qui a la plus forte probabilité d'avoir généré la séquence des observations ?

- Apprentissage : soient un modèle $\Phi$ et un ensemble d'observations $X$.

➔ Comment ajuster les paramètres du modèle $\Phi$ pour maximiser la probabilité $P(X \mid \Phi)$ ?

Le problème de l'évaluation est résolu par l'algorithme *Forward*. Le problème du décodage peut être résolu en utilisant l'algorithme de *Viterbi*. Enfin, le problème d'apprentissage du modèle peut être résolu par l'algorithme *Baum-Welch* (ou *Forward-Backward*). On trouve de plus amples informations sur les modèles de Markov cachés et ces algorithmes dans [Rabiner, 1993].

### *1.2.4.3. Modélisation acoustique indépendante ou dépendante du contexte*

Pour la modélisation acoustique à base des modèles de Markov cachés, les séquences de mots sont divisées en unités de base, fréquemment les phonèmes.

Dans la modélisation acoustique indépendante du contexte, un phonème est modélisé généralement par un seul HMM à trois états. Par exemple, pour modéliser les 43 phonèmes du français [Lamel, 1991], nous avons besoin de 43 HMMs et le nombre total d'états est alors 129 seulement.

Mais, il est connu depuis longtemps que la performance d'un système de reconnaissance de la parole native est plus élevée si les phonèmes sont modélisés de façon dépendante du contexte. Par exemple, un /a/ précédé d'un /ɲ/ n'est pas identique à un /a/ précédé d'un /m/. Cela est la conséquence du phénomène bien connu de coarticulation ou d'anticipation des gestes de production de la parole, d'où l'utilisation de polyphones, où chaque phonème est caractérisé par son contexte précédent et/ou suivant.

Pour l'utilisation de monophones, on parle de modélisation acoustique indépendante du contexte. Pour l'utilisation de polyphones, on parle de modélisation acoustique dépendante du contexte. Bien entendu, la modélisation acoustique dépendante du contexte engendre une très importante quantité d'unités.

Dans les conditions réelles où le nombre de représentants de polyphones dans le corpus d'apprentissage est insuffisant, nous ne pouvons pas modéliser tous les contextes possibles des phonèmes. Pour résoudre ce problème, [Huang, 2001] a proposé une approche nommée « *clustering* » sur un arbre de décision, illustrée par la figure 1.5. Cette approche consiste à construire des groupes de polyphones de contexte similaire en appliquant une question de contexte du type, par exemple, « Le contexte gauche est-il une voyelle ? » « Le contexte droit est-il le phonème /ɲ/ ? », etc.

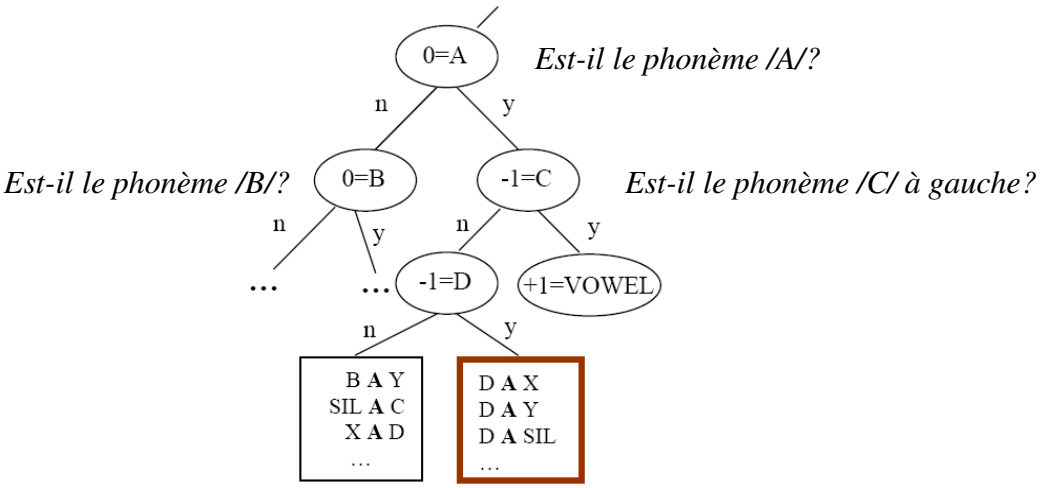



*Figure 1.5 : Exemple d'arbre de décision proposé par [Huang, 2001]*

### 1.2.5. Dictionnaire de prononciation

Le dictionnaire de prononciation (ou dictionnaire phonétique) fournit le lien entre les séquences des unités acoustiques et les mots représentés dans le modèle de langage.

Il est important de noter que les performances du système de reconnaissance sont directement liées au taux de mots hors vocabulaire. Bien qu'un dictionnaire de prononciation créé manuellement permette une bonne performance, la tâche est très lourde à réaliser et demande des connaissances approfondies sur la langue en question. En plus, les noms propres sont l'un des problèmes majeurs pour toutes les langues. Par exemple, les 20 000 noms propres inclus dans le dictionnaire anglais *COMPLEX* ne représentent qu'une petite fraction des un à deux millions de noms rassemblés par [Huang, 2001] sur des données en anglais US.

Pour résoudre ces problèmes, la littérature propose des approches qui permettent de générer automatiquement le dictionnaire de prononciation. Une des approches de la génération automatique d'un dictionnaire de prononciation consiste à utiliser des règles de conversion graphème-phonème. Cette construction nécessite une bonne connaissance de la langue et de ses règles de phonétisation, qui par ailleurs ne doivent pas contenir trop d'exceptions. Cette approche est donc applicable aux langues avec des prononciations assez régulières comme l'espagnol et l'italien. Pour l'anglais qui est très varié au niveau de la prononciation, l'approche automatique à base de règles n'est pas recommandée. Dans ce cas, les prononciations des mots hors vocabulaire d'une langue peuvent être générées en utilisant le décodage acoustico-phonétique de cette langue.

Alternativement, l'approche, simple et totalement automatique, qui utilise des graphèmes comme unités de modélisation (dictionnaire de prononciation à base de graphèmes) a été validée dans [Billa, 2002] et [Bisani, 2003].

### 1.2.6. Modélisation du langage

Un système de reconnaissance automatique de la parole continue à grand vocabulaire dépend généralement fortement de la connaissance linguistique de la parole. Les meilleurs systèmes de décodage acoustico-phonétique qui n'utilisent aucun modèle de langage n'atteignent qu'un taux d'exactitude en phonèmes de l'ordre de 50 % environ. La modélisation du langage est donc une réelle nécessité pour la reconnaissance automatique de la parole continue à grand vocabulaire. Un module linguistique est nécessaire dans le système pour déterminer la forme lexicale correspondante, c'est-à-dire la séquence de mots la plus probable, au sens langagier.

Dans l'équation bayésienne appliquée à la reconnaissance automatique de la parole (équation 1.1 précédente) apparaît une probabilité a priori de la séquence. Cette probabilité se calcule à partir d'un modèle de langage. Ainsi, la séquence « je suis ici » est plus probable, en terme de langage, que « jeu suis ici », ou encore « jeux suit y si », bien que l'acoustique soit quasi-similaire. Pour une même suite de phonèmes, il peut exister plusieurs dizaines de phrases possibles. Le rôle principal du modèle de langage est de les classer selon leur plausibilité linguistique.

Dans la section suivante, nous présentons les modèles n-grammes qui sont les modèles les plus utilisés pour la modélisation du langage.

#### *1.2.6.1. Les modèles n-grammes*

Un modèle statistique du langage consiste, pour une séquence de mots $M = m_1, m_2, \ldots, m_N$, à calculer la probabilité $P(M)$ :

$$P(M) = \prod_{i=1}^{N} P(m_i \mid m_1, \ldots, m_{i-1}) = \prod_{i=1}^{N} P(m_i \mid h_i) \qquad (1.4)$$

où $h_i = m_1, \ldots, m_{i-1}$ est considéré comme l'historique du mot $m_i$ et $P(m_i|h_i)$ est la probabilité du mot $m_i$, connaissant tous les mots précédents.

En pratique, au fur et à mesure que la séquence de mots hi s'enrichit, une estimation des valeurs des probabilités conditionnelles $P(m_i|h_i)$ devient de plus en plus difficile car aucun corpus de texte d'apprentissage ne peut permettre d'observer toutes les combinaisons possibles de $h_i = m_1, \ldots, m_{i-1}$.

Afin de réduire la complexité du modèle de langage, et par conséquent de son apprentissage, l'approche n-gramme peut être utilisée. Le principe est le même, mais l'historique est limité aux n-1 mots précédents (hypothèse de Markov). La probabilité P(M) est donc approchée par :

$$P(M) = \prod_{i=1}^{N} P(m_i \mid m_{i-n+1}, ..., m_{i-1}) \tag{1.5}$$

Les modèles de langage n-grammes sont assez souples car ils permettent de modéliser des phrases grammaticalement incorrectes mais ils n'interdisent pas non plus de produire des phrases incorrectes syntaxiquement. On parle alors de modèle unigramme si n = 1 (sans historique), bigramme si n = 2 ou trigramme si n = 3, etc. Les modèles les plus couramment utilisés en RAP sont les modèles d'ordre trois à cinq.

### *1.2.6.2. Estimation des modèles de langage*

L'estimation des paramètres d'un modèle de langage à n-grammes s'effectue en deux opérations : une opération de décompte et une opération de redistribution des probabilités. La méthode d'estimation effectue un décompte des suites de mots observés afin d'en extraire une probabilité d'apparition. Le principe est d'estimer toutes les probabilités issues des événements observés, puis de les redistribuer à des événements non vus. Cette seconde étape, qui correspond au lissage, permet d'associer une probabilité non nulle à des événements jamais observés sur le corpus d'apprentissage. Les méthodes de lissage classiques calculent une probabilité non nulle en réduisant la fenêtre d'observation.

Les modèles à n-grammes sont donc très dépendants du corpus d'apprentissage, et ont un champ de vision limité à la taille n du n-gramme (qui est comprise entre trois et cinq généralement). Même pour les langues bien dotées, les quantités de texte disponibles pour estimer les probabilités des n-grammes ne sont pas suffisantes pour les n-grammes d'ordre plus élevé. De nombreuses techniques de lissage ont été proposées pour pallier ce problème. L'une des techniques de lissage les plus utilisées est la technique dite de Kneser-Ney [Kneser, 1995]. Avec cette technique, les probabilités des n-grammes peu observés sont estimées comme avec les autres techniques de lissage, en faisant un repliement (*backoff*) sur un historique d'ordre moins grand. Pour un trigramme, par exemple, le bigramme puis l'unigramme si nécessaire sont utilisés. L'originalité de la technique Kneser-Ney modifiée est de ne pas prendre la même distribution de probabilités pour les ordres plus petits que n. Au lieu de prendre la fréquence de l'historique d'ordre $n-1$, à savoir $h_{i-n+1}^{i-1}$, c'est le nombre de contextes différents dans lesquels se produit $h_{i-n+1}^{i-1}$ qui est utilisé. L'idée est que, si ce nombre

est faible, alors la probabilité accordée au modèle d'ordre $n-1$ doit être petite et ce, même si $h_{i-n+1}^{i-1}$ est fréquent. Ainsi, le biais potentiel introduit par la fréquence de l'historique est évité.

Les modèles à n-grammes sont extrêmement simples, mais ont prouvé leur efficacité et leur souplesse. Ils se sont imposés dans les systèmes dits « état de l'art », bien que diverses alternatives efficaces aient été proposées dans la littérature (par exemple, modèles continus de [Schwenk, 2002] et [Schwenk, 2007]); ils continuent d'être quasi-systématiquement intégrés aux systèmes de RAP à l'état de l'art.

#### *1.2.6.3. Évaluation des modèles de langage*

La qualité d'un modèle de langage dépend de sa capacité à influencer le système de reconnaissance automatique de la parole afin d'en augmenter la performance. Une question primordiale est de savoir comment deux modèles de langage peuvent être comparés en terme de performances dans un système de reconnaissance. La façon correcte de procéder consiste à intégrer chaque modèle dans un système complet, et à évaluer quelle est la meilleure transcription en sortie du système. Cette méthode permet d'évaluer concrètement la performance d'un modèle de langage, mais nécessite de disposer d'un système complet.

La mesure la plus couramment utilisée consiste à estimer la perplexité de chacun des modèles. La perplexité d'un modèle de langage correspond à sa capacité de prédiction. Plus la valeur de la perplexité est petite, plus le modèle de langage possède des capacités de prédiction. La perplexité s'estime sur le corpus d'apprentissage pour définir si les modèles choisis modélisent correctement le corpus. Elle est calculée sur un corpus de test ou de développement, pour estimer le degré de généralisation du modèle. Cependant, bien que la perplexité permette d'estimer la capacité de représentation d'un modèle de langage, elle n'est pas systématiquement corrélée avec la qualité du décodage. Pour des modèles n-grammes, la perplexité (*PP*) se définit ainsi :

$$PP(M) = 2^{-\frac{1}{n}\sum_{t=1}^{n}\log_2 P(m_t|h)} \tag{1.6}$$

où $P(m_t|h)$ est la probabilité associée au n-gramme $(m_t|h)$.

Deux remarques importantes sont à prendre en considération lorsque l'on compare des modèles de langage :

- une réduction de perplexité n'implique pas toujours un gain de performances pour un système de reconnaissance ;

- en général, la perplexité de deux modèles n'est comparable que s'ils utilisent le même vocabulaire ; sinon, il faut utiliser une perplexité normalisée qui simule un nombre de mots identique.

Bien que des modèles de langage avec des mesures de perplexité qui diminuent tendent à améliorer les performances d'un système de reconnaissance, on trouve dans la littérature des études qui décrivent des cas où des diminutions importantes de perplexité ont peu ou pas apporté de gain de performance [Lyer, 1997; Martin, 1997b].

### 1.2.7. Les outils de développement

Aujourd'hui, il existe beaucoup d'outils de développement pour les systèmes de reconnaissance automatique de la parole. HTK [Young, 1994] et Sphinx [Lee, 2002] sont les deux logiciels les plus connus et les plus utilisés dans le domaine. Ils sont téléchargeables gratuitement, avec la possibilité de mettre en œuvre des systèmes de reconnaissance à grand vocabulaire, indépendants du locuteur, et traitant de parole continue dans n'importe quelle langue. Bien que HTK ait été utilisé par divers groupes pendant une dizaine d'années, Sphinx, qui est toujours en cours de développement dans un environnement universitaire, est aussi intéressant, en raison de certaines fonctionnalités avancées et de sa licence d'utilisation sans restriction. Ces deux logiciels ont été comparés sur la mise en œuvre d'un système de reconnaissance vocale des hindi. Bien que les précisions de reconnaissance des deux systèmes soient comparables, on constate que la modélisation acoustique de Sphinx est supérieure. Le lecteur peut trouver les comparaisons entre les deux logiciels dans [Samudravijaya, 2003].

Sphinx est utilisé dans l'équipe GETALP du laboratoire LIG comme l'outil de base pour les recherches en reconnaissance de la parole. Les boîtes à outils de Sphinx principalement utilisées sont :

- SphinxTrain pour entraîner des modèles acoustiques ;
- Sphinx3 pour décoder la parole en texte.

Pour construire le modèle de langage, nous utilisons l'outil SRILM [Stolcke, 2002]. Il existe cependant d'autres librairies similaires, telles que CMU-SLM (*Carnegie Mellon Statistical Language Modeling toolkit*)[1].

Au niveau des outils d'évaluation de la performance d'un système RAP, la boîte à outils SCTK[2] (*Scoring Toolkit*) du *National Institute of Standards and Technologies* (NIST) fournit

[1] http://www.speech.cs.cmu.edu/SLM_info.html
[2] http://www.nist.gov/speech/tools

le programme « *sclite* » pour aligner les hypothèses et les références, calculer le taux d'erreur (WER), et faire des analyses fines des erreurs. Cet outil peut fournir des informations très utiles comme les mots les plus substitués, insérés ou supprimés. Des taux d'erreur par locuteur peuvent être également obtenus (si les segments possèdent une étiquette de locuteur).

## 1.3. Reconnaissance acoustico-phonétique multilingue

La plupart des techniques de reconnaissance automatique de la parole, notamment les systèmes de reconnaissance automatique de la parole continue à grand vocabulaire, utilisent des aujourd'hui approches statistiques. Cependant, la nature statistique des approches nécessite de disposer d'un grand nombre de données (textuelles et signaux) pour entraîner les modèles sous-jacents et tester les performances des systèmes. Par conséquent, un grand corpus de parole qui contient des dizaines d'heures de signal enregistrées par une centaine de locuteurs (pour la modélisation acoustique) et un corpus de texte propre avec des millions de mots écrits (pour la modélisation statistique du langage) sont nécessaires pour le développement d'un système de reconnaissance automatique de la parole continue à grand vocabulaire. Ces ressources ne sont, bien sûr, pas disponibles directement pour des langues moyennement ou peu dotées.

Avec l'émergence de la reconnaissance automatique de la parole multilingue [Schultz, 2006], plusieurs solutions à base de modèles multilingues (les modélisations acoustique, lexicale et du langage) sont proposées pour résoudre plusieurs problèmes, notamment dans le domaine de la reconnaissance automatique de langues non encore traitées et qui n'ont pas suffisamment de ressources pour amorcer les modèles (les langues peu dotées), l'identification de la langue, la reconnaissance de la parole non native, etc.

Comme nos enjeux de recherche sont centrés sur l'adaptation des modèles acoustiques multilingues, nous présentons dans cette section les méthodes de combinaison des modèles acoustiques pour créer un modèle multilingue. Ensuite, nous résumons quelle utilisation est faite des modèles acoustiques multilingues en traitement automatique de la parole.

### 1.3.1. Méthode de combinaison des modèles acoustiques

Les principaux objectifs de la combinaison des modèles acoustiques sont la réduction du nombre global des paramètres du modèle acoustique, et l'amélioration de la robustesse des modèles acoustiques.

T. Schultz introduit trois méthodes différentes pour combiner des modèles acoustiques [Schultz, 2001] : ML-sep (séparation des langues), ML-mix (mélange des langues) et ML-tag (étiquetage des langues).

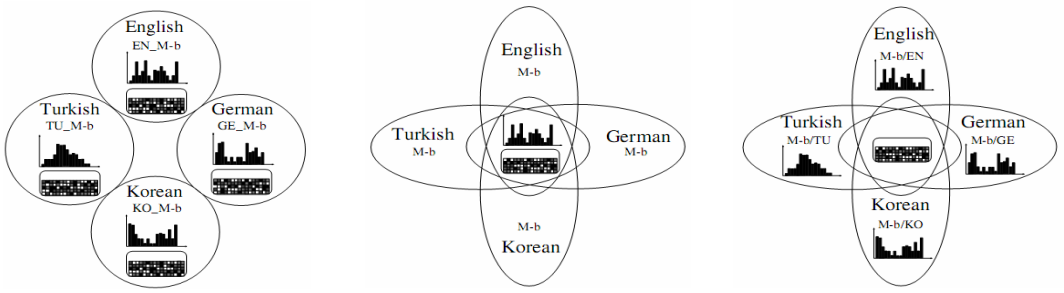



*Figure 1.6 : Combinaison de modèles acoustiques ML-sep, ML-mix, ML-tag (de gauche à droite) [Schultz, 2001]*

Dans le modèle HMM continu (section 1.2.4.2), la probabilité $P(x \mid e_i)$ d'émettre $x$ à l'état $e_i$ est définie par $P(x \mid e_i) = \sum_{g=1}^{G_i} c_{e_i g} . N(x \mid \mu_{e_i g}, \Sigma_{e_i g})$. La figure 1.6 illustre les trois différentes méthodes de combinaison des modèles acoustiques. Elle décrit la combinaison de modèles acoustiques pour le début d'état du modèle du phonème /m/, donc « M-b ». Dans chaque état d'une de ces trois sous-figures, les distributions multi-gaussiennes (*distribution weights*) $c$ sont symbolisées par un histogramme et les composants gaussiens (*codebook weights*) $N(x \mid \mu, \Sigma)$ sont symbolisés par une ellipse avec des petits points colorés (blanc, gris, noir).

Pour la méthode ML-sep, chaque modèle du phonème dans une langue est appris uniquement sur les données correspondantes de cette langue : il n'y a pas de données à partager entre les langues. Selon les probabilités d'émission mentionnées au-dessus, la méthode de combinaison ML-sep peut être décrite simplement comme suit :

$$\text{ML-sep:} \begin{cases} c_{e_i} \neq c_{e_j} & , \quad \forall i \neq j \\ \mu_{e_i,k} \neq \mu_{e_j,k} & , \quad \forall i \neq j \\ \Sigma_{e_i,k} \neq \Sigma_{e_j,k} & , \quad \forall i \neq j \end{cases}$$

Pour la méthode ML-mix, les données sont partagées entre les langues pour entraîner des modèles communs. Cela signifie que les composants gaussiens (*codebook weights*) et les distributions multi-gaussiennes (*distribution weights*) des modèles d'un phonème sont partagés entre les langues :

$$\text{ML-mix:} \begin{cases} c_{e_i} = c_{e_j} & , \quad \forall i, j : \text{ipa}(e_i)=\text{ipa}(e_j) \\ \mu_{e_i,k} = \mu_{e_j,k} & , \quad \forall i, j : \text{ipa}(e_i)=\text{ipa}(e_j) \\ \Sigma_{e_i,k} = \Sigma_{e_j,k} & , \quad \forall i.j : \text{ipa}(e_i)=\text{ipa}(e_j) \end{cases}$$

La méthode ML-tag est une méthode hybride où chaque modèle du phonème appartenant à une langue est associée par une étiquette de langue, afin de préserver l'information de cette langue. Pour la méthode ML-tag, les composants gaussiens (*codebook weights*) des modèles du phonème sont partagés entre les langues (comme dans la méthode ML-mix), tandis que les distributions multigaussiennes (*distribution weights*) des modèles sont apprises séparément (comme la méthode ML-sep) :

$$\text{ML-tag:} \begin{cases} c_{e_i} \neq c_{e_j} & , \quad \forall i \neq j \\ \mu_{e_i,k} = \mu_{e_j,k} & , \quad \forall i, j : \text{ipa}(e_i)=\text{ipa}(e_j) \\ \Sigma_{e_i,k} = \Sigma_{e_j,k} & , \quad \forall i.j : \text{ipa}(e_i)=\text{ipa}(e_j) \end{cases}$$

### 1.3.2. Différentes utilisations de la modélisation acoustique multilingue

#### *1.3.2.1. Portabilité des modèles acoustiques vers une nouvelle langue*

Pour créer le modèle acoustique d'une nouvelle langue, si une grande quantité de données est disponible (plusieurs dizaines ou centaines d'heures par exemple), la création du modèle acoustique peut correspondre à un simple réapprentissage des modèles sur ces nouvelles données.

Dans le contexte des langues peu dotées, la quantité de données collectées reste bien souvent inférieure à ce qu'elle est pour les langues bien dotées. La construction du modèle acoustique nécessite donc également des techniques d'adaptation rapide en utilisant les modèles acoustiques multilingues génériques qui couvrent un grand nombre de phonèmes comme ceux proposés dans [Schultz, 2002] par exemple. Lors de la phase d'adaptation, chaque segment de parole est aligné soit de manière itérative avec l'algorithme de *Baum-Welch* [Baum, 1970], qui prend en compte tous les chemins qui passent par un état, soit uniquement avec la meilleure séquence d'états possible (alignement de type *Viterbi*). Après l'alignement, les paramètres des HMM peuvent être adaptés par une technique de type MAP [Gauvain, 2002]. Cette technique a été utilisée avec succès dans [Le, 2006] pour le développement d'un système de RAP pour la langue vietnamienne.

### *1.3.2.2. Prise en compte du « code-switching »*

Le code-switching (ou alternance codique) est lié à l'utilisation spontanée, dans la parole d'une langue principale ($L_{princ}$), de mots, phrases ou expressions d'une autre langue ($L_{autre}$). Par exemple, le mot *email* de l'anglais est utilisé dans plus de 12 sur 100 conversations téléphoniques dans le corpus mandarin *CallHome*[3].

La plupart des systèmes de RAP ne possèdent pas de mécanisme pour modéliser le code-switching. Les erreurs sont attribuées aux mots « hors langue » $L_{princ}$ (*out-of-language,* OOL). Ces mots hors vocabulaires sont souvent riches en information, et les reconnaître est particulièrement important dans des applications telles que la détection des mots clés (*key-term detection*).

Concernant ce problème, deux approches pour la modélisation acoustique ont été examinées par [White, 2008]. La première approche consiste à créer les prononciations de mots hors vocabulaire dans le lexique et à utiliser les mêmes modèles acoustiques de $L_{princ}$ ; la deuxième approche consiste à utiliser les prononciations hors vocabulaires décrites dans le système phonétique $L_{autre}$ et à utiliser des modèles acoustiques multilingues. L'article référencé ci-dessus montre que cette dernière approche (utilisation des modèles acoustiques multilingues) est meilleure pour traiter les mots hors vocabulaires.

### *1.3.2.3. Identification automatique de la langue parlée (LID)*

L'identification automatique de la langue parlée (LID) est le processus qui consiste à déterminer la langue correspondant à un ensemble donné de parole. C'est une technologie clé dans de nombreuses applications telles que les systèmes de conversation multilingue [Zue, 2000], la traduction de parole [Waibel, 2000], la reconnaissance de la parole multilingue [Ma, 2002], et la recherche de documents vocaux [Dai, 2003]. La LID est également l'un des enjeux très importants dans les domaines du renseignement et de la sécurité, où l'étiquetage en langues des messages enregistrés doit être établi avant qu'une quelconque information ne puisse en être extraite.

Une langue parlée peut être identifiée en utilisant des informations provenant de sources multiples. Lors des dernières décennies, les chercheurs ont exploré de nombreuses caractéristiques de la parole, comprenant les paramètres articulatoires [Kirchhoff, 2002], les caractéristiques acoustiques [Sugiyama, 1991], la prosodie [Adami, 2003; Adda-Decker, 2003], la phonotactique [Hazen, 1993; Zissman, 1996a], et les connaissances lexicales

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96S34

[Matrouf, 1998]. Profitant des progrès récents en reconnaissance de la parole continue [Gauvain, 2000], des chercheurs du laboratoire *Lincoln Labs* du MIT (*Massachusetts Institute of Technology*) ont présenté des résultats prometteurs en utilisant des caractéristiques acoustiques appelées *shifted-delta-cepstral* [Torres-Carrasquillo, 2002] dans le modèle de distribution multi-gaussiennes (GMM), qui peut être considéré comme un modèle de Markov à un seul état [Rabiner, 1989]. Une autre approche efficace consiste à caractériser une langue parlée en utilisant des distributions de probabilités des caractéristiques spectrales sous la forme d'unités symboliques telles que des phonèmes ou des unités syllabiques [Hazen, 1993; Nagarajan, 2004; Zissman, 1996a], où les modèles de phonèmes sont utilisés pour convertir la parole en séquences de symboles phonétiques, accompagnés de scores de vraisemblance acoustique. Des modèles n-grammes sont ensuite construits pour chaque langue afin de produire des scores phonotactiques de vraisemblance. Le tableau 1.3 présente une comparaison de quelques approches standard à la LID.

L'approche phonotactique a montré qu'elle pouvait fournir des performances supérieures dans la campagne d'évaluation de la reconnaissance de la langue du NIST (*National Institute of Standards and Technology*) surtout quand elle est fusionnée avec des scores acoustiques [Singer, 2003].

[Zissman, 1996a] a proposé deux approches phonotactiques très utilisées dans le domaine, appelées PPR-LM pour *Parallel Phone Recognizers followed by n-gram Language Models* et UPR-LM pour *Universal Phone Recognizers followed by n-gramme Language Models*. La première utilise plusieurs décodeurs acoustico-phonétiques en entrée (*frontend*) et des modèles n-grammes (modèles de langage) pour la décision (*backend*). La deuxième utilise un seul décodeur acoustico-phonétique (monolingue ou multilingue) en entrée, tandis que le *backend* est similaire à celui de l'approche précédente. La figure 1.7 et la figure l.8 illustrent l'architecture de l'approche phonotactique PPR-LM et UPR-LM respectivement.

En 2007, [Li, 2007] a proposé un nouveau protocole pour la partie postérieure des deux approches précédentes. Ces nouvelles approches sont basées sur la technique VSM (*Vector Space Modeling*) très connue dans le domaine de la recherche d'information. Cette technique utilise en fait des SVM (*Support Vector Machine*), outil classique en reconnaissance des formes, pour identifier la langue parlée. Cette technique est présentée plus en détail (UPR-VSM) dans le chapitre 3 du manuscrit, puisque nous l'utiliserons lors de nos recherches.

| Approche | Points forts | Limitation | Application |
|---|---|---|---|
| Prosodie | robustesse au changement de canal | utilisation appropriée pour distinguer des groupes de langues | pré-classification des tons d'une langue en deux différents groupes |
| Acoustique | coût faible dans l'entraînement et le test (données et calcul) | elle est utilisée en combinaison avec d'autres composants | identification de langue et de l'accent dans un système multi-reconnaisseurs |
| Phonotactique | bon rapport de performance aux coûts de développement en comparant avec ceux des autres approches, pas besoin de connaissances linguistiques pour entraîner le modèle | elle est très utile pour les segments de tests avec une durée > 5 secondes | système LID avec grande quantité de langues, y compris des langues rares, sans données étiquetées pour l'entraînement |

*Tableau 1.3 : Comparaison entre approches standard LID du point de vue de l'aspect du développement applicatif*

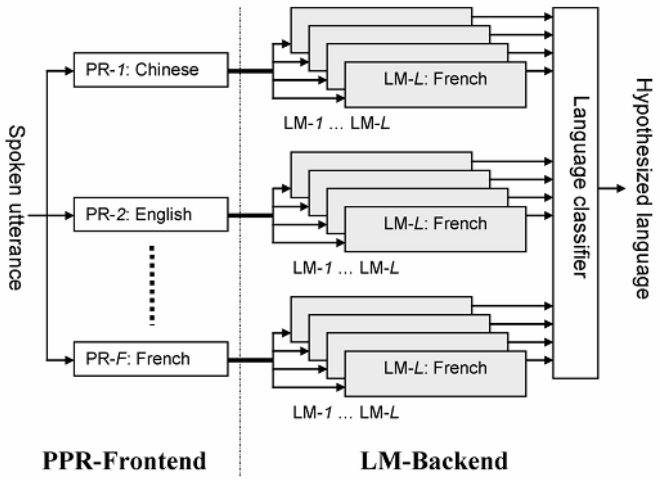



*Figure 1.7 : Architecture de l'approche phonotactique PPR-LM [Zissman, 1996a]*

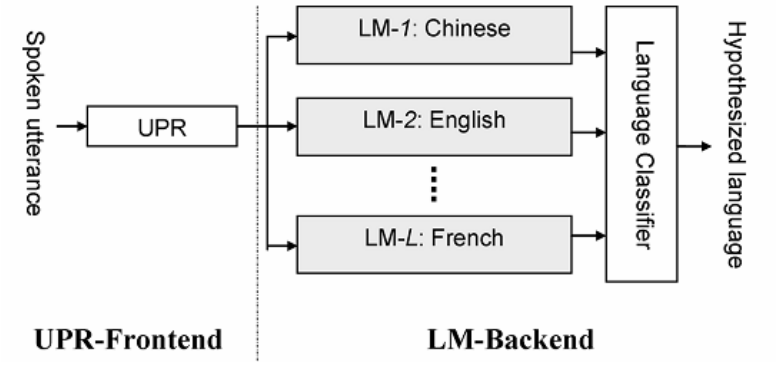



*Figure 1.8 : Architecture de l'approche phonotactique UPR-LM [Zissman, 1996a]*

Dans toutes les approches mentionnées dans le paragraphe précédent, il est montré que les approches UPR-VSM et UPR-LM qui utilisent des décodeurs acoustico-phonétiques multilingues en entrée donnent une performance comparable avec celles des approches PPR-VSM et PPR-LM. Cela met en évidence que les modèles acoustiques multilingues sont utiles dans la tâche d'identification automatique de la langue parlée (LID).

### *1.3.2.4. Recherche de documents multilingues : le cas du Star-Challenge*[4]

Le *Star-Challenge* est une compétition visant à encourager le développement de systèmes de recherche par le contenu dans les documents vidéo. Les systèmes sont évalués dans un cadre très général incluant des modalités de recherche basées sur le son, sur l'image ou sur une combinaison des deux. Par ailleurs, les documents à traiter sont dans des langues multiples et inconnues.

Concernant le traitement de la bande son des vidéos, la difficulté de *Star-Challenge* réside dans le fait que les documents sont de nature multilingue, sans que la langue parlée sur les documents soit connue à l'avance. Comme il est impossible d'appliquer en parallèle des systèmes de reconnaissance pour toutes les langues envisagées (contraintes liées au temps de développement et au temps de calcul pour le traitement des documents et des requêtes), l'approche conduite par le LIG, lors de sa participation au challenge [Quénot, 2010], a consisté à construire un système acoustico-phonétique multilingue (représentant, en quelque sorte, une bonne couverture des sons des langues du monde), afin de transcrire phonétiquement les requêtes et les documents de la base. La recherche se fait donc par appariement de séquences de symboles phonétiques, sans qu'aucune transcription en mots ne soit réalisée. Dans le cadre du *Star-Challenge*, le modèle acoustique multilingue inclut plus particulièrement les sons des langues anglaise, chinoise et malaise, afin de traiter plus spécifiquement des vidéos issues de la région de Singapour.

[4] http://www.thestarchallenge.sg/

### 1.3.3. Transcription automatique des langues en danger

Dans le cadre du projet PI[5], le LIG effectue des études linguistiques et développe des outils informatiques pour sauvegarder une langue parlée par une minorité, le mo piu. Cette minorité se situe dans les montagnes au nord du Vietnam. Selon des études réalisées à MICA en 2010, le mo piu est vraiment une langue en danger, car elle est non répertoriée, non documentée, sans écriture (langue orale) et parlée par seulement 227 personnes. D'ailleurs, aucune ethnie environnante ne comprend cette langue. Dans une étude à laquelle j'ai participé, nous avons utilisé des modèles acoustiques multilingues pour transcrire les paroles de mo piu en séquences phonétiques. L'objectif était de savoir si les transcriptions automatiques fournies par les systèmes acoustico-phonétiques multilingues peuvent faciliter les travaux des phonéticiens dans leurs analyses des unités acoustiques et phonétiques de la parole pour le mo piu. Notre étude préliminaires sur la transcription automatique du mo piu effectuée par différents systèmes acoustico-phonétiques multilingues, ainsi que les protocoles d'analyse acoustique, phonétique et les résultats d'expérimentation se trouve dans l'annexe C de ce mémoire.

### 1.3.4. Utilisation de la modélisation acoustique multilingue pour traiter la parole non native

La variabilité acoustique à travers les locuteurs est un problème commun à de multiples tâches de traitement automatique de la parole. Par exemple, lorsqu'on considère la parole non native et les dialectes, les variations peuvent devenir trop importantes pour être traitées en utilisant des systèmes de RAP traditionnels appris sur de la parole native de la langue considérée.

Nous trouvons dans la littérature plusieurs études consistant à créer ou adapter les modèles acoustiques pour la parole non native.

- Uebler a étudié la tâche de reconnaissance de la parole allemande de différents accents [Uebler, 1999] ; il a constaté que la mise en commun des données de la parole native et de la parole non native (*pooling approach*) pour construire un modèle acoustique global donne une performance assez uniforme pour des locuteurs non natifs ; toutefois, d'autres résultats suggèrent que les modèles acoustiques des accents spécifiques donnent de meilleures performances.

[5] Le projet PI (ANR BLANC 2009-2010, http://pi.imag.fr) concerne le traitement automatique du langage parlé (notamment la reconnaissance automatique de la parole) pour les langues peu dotées ou pas dotées. Les partenaires sont le LIG, le LIA, et le centre international MICA (Hanoï, Vietnam).

- [Livescu, 1999] entraîne les modèles acoustiques multilingues en mettant en commun des données des langues parlées par les locuteurs de différents dialectes (natifs et non natifs). Il applique ces modèles acoustiques à la reconnaissance vocale non native et trouve que ces modèles sont plus performants que les modèles acoustiques qui avaient été entraînés uniquement sur la parole soit native soit non native.
- Les deux études précédentes nécessitent beaucoup de données non natives pour entraîner les modèles, ce qui rend difficile leur application à une grande variété de dialectes et de langues dans le monde (problème du recueil de données non natives). Actuellement, les approches très répandues pour les systèmes RAPs non natifs consistent à appliquer des techniques d'adaptation des modèles acoustiques natifs en utilisant très peu de données non natives pour construire les modèles acoustiques adaptés. [Wang, 2003] a comparé quatre approches d'adaptation des modèles acoustiques multilingues en appliquant très peu de données non natives aux modèles natifs existant. [Tan, 2007] a proposé une approche hybride d'interpolation (interpolation et fusion) des modèles acoustiques des langues maternelles des locuteurs (L1) avec ceux des langues parlées (L2). Cette dernière approche ne nécessite aucune donnée non native dans le processus d'adaptation des modèles acoustiques multilingues. Nous donnons plus de détails sur ces approches d'adaptation des modèles acoustiques multilingues dans la section suivante.

En résumé, les modèles acoustiques multilingues sont largement utilisés, non seulement dans le domaine de la reconnaissance de la parole native ou non native, mais aussi en identification de la langue parlée (LID), en suivi du code-switching, et en recherche d'information multilingue.

## 1.4. Reconnaissance de la parole pour les locuteurs non natifs

### 1.4.1. Acquisition de la langue

Les études sur l'acquisition du langage peuvent être divisées en : acquisition de la langue première (ou langue maternelle, L1) et acquisition d'une langue seconde (ou langue étrangère, L2). L'acquisition de L1 concerne le développement de la langue première (maternelle) chez les enfants, tandis que l'acquisition d'une langue seconde considère le processus d'apprentissage d'une nouvelle langue différente de la langue maternelle. L1 et L2

sont acquises de manière différente et souvent à différents stades de la vie, ce qui peut affecter la capacité linguistique des locuteurs. Avant d'étudier le système de reconnaissance de la parole non native, il est important de comprendre d'abord comment se déroule le développement de L1 chez les nouveaux-nés, comparé aux adultes qui apprennent une nouvelle langue, et comment ces différences affecteront le système de RAP.

### *1.4.1.1. Acquisition de L1*

Pendant des années, des chercheurs se sont intéressés à la façon dont les enfants peuvent maîtrisé la langue L1, sans effort, au bout d'une courte période. La compréhension de cette capacité d'apprentissage peut aider les chercheurs à proposer une meilleure approche pour apprendre une deuxième langue. Selon les études de [O'Grady, 2000], la capacité des enfants à distinguer des phonèmes est bien développée avant leur capacité à produire de la parole. Les nouveau-nés peuvent discriminer les unités phonétiques de toutes les langues grâce au mécanisme de traitement auditif général (celui-ci existe aussi chez quelques animaux tels que les singes). Par exemple, selon une étude[6] du cerveau humain publiée dans le journal scientifique *ScienceDaily*, les caractéristiques auditives des locuteurs des langues tonales activent des régions différentes du cerveau par rapport à celles des langues non-tonales. Les nouveaux-nés commencent le babillage à partir d'environ six mois. Les études prouvent qu'il y a des similitudes parmi les sons de babillage des enfants, bien qu'ils soient d'origine différente. Parmi ces sons, les consonnes plosives sont plus fréquentes, comparées aux fricatives [O'Grady, 2000]. Les mots compréhensibles sont prononcés chez les enfants d'environ douze mois, avec des expressions unitermes. La suppression de syllabes est commune dans les mots prononcés. Il y a également des suppressions systématiques de certains sons pour simplifier la syllabe, par exemple le mot '*stop'* est prononcé comme le mot '*top'*. Des phrases plus complexes et plus longues sont exprimées plusieurs mois après cela.

### *1.4.1.2. Acquisition de L2*

Il est bien connu que l'âge pour apprendre une langue joue un rôle important pour déterminer à quel point on maîtrisera cette langue. [Kim, 1997] prouve que les jeunes bilingues activent une région commune du secteur *Brodmann* dans le cerveau, tandis que les sujets bilingues qui acquièrent une langue à un âge plus élevé activent deux régions distinctes de cerveau pour traiter les deux langues. Les études suggèrent qu'il y a une période critique

[6] http://www.sciencedaily.com/releases/2008/02/080216114856.htm

pour apprendre la langue L1 et également la langue L2. Après cette période, la capacité à acquérir une langue L2 avec succès diminuera.

Pour l'acquisition de L2, quelques travaux prouvent qu'il y a un rapport linéaire plutôt qu'un seuil entre l'âge de l'étude et l'accent perçu [Flege, 1995]. L'imprécision de la perception est évoquée comme l'une des raisons principales pour lesquelles les locuteurs non natifs ne peuvent pas articuler comme les locuteurs natifs [Flege, 1995; Kuhl, 2000; Rochet, 1995]. Une autre recherche montre que l'utilisation fréquente de la langue maternelle L1 par une personne peut également augmenter l'accent perçu en L2, même si la personne apprend la langue L2 depuis son enfance [Flege, 1997; Flege, 2004].

### *1.4.1.3. Erreur d'acquisition de L2*

Les erreurs d'acquisition d'une langue seconde L2 impliquent quatre niveaux linguistiques [Flege, 1995; O'Grady, 2000] : phonologie, prononciation, vocabulaire et grammaire.

Le tableau 1.4 présente une brève description des erreurs d'acquisition de langue L2 par les locuteurs non natifs. Il montre que la plupart des erreurs (pour les quatre niveaux) produites par les locuteurs non natifs sont liées à l'impact de la langue maternelle (L1) des locuteurs non natifs sur l'acquisition de L2.

| **Type d'erreur** | **Raisons principales** |
|---|---|
| Phonologie | Transfert des phonèmes et de la prosodie de L2 en langue L1.<br>Ex: des locuteurs anglais prononcent le phonème /y/ français comme /u/. |
| Prononciation | Simplification/Modification de prononciation des mots en langue L2 en se basant sur L1.<br>Ex : des locuteurs espagnols prononcent le mot '*spécial*' comme '*especial*'. |
| Vocabulaire | Utilisation du mot L2 comme L1.<br>Ex : les locuteurs italiens traduisent souvent le mot '*embarrassed*' (anglais) comme '*embarasado*' (italien). |
| Grammaire | Utilisation de la grammaire de L1 dans L2<br>Ex : on peut trouver la phrase '*Marie watches <u>often</u> television'* chez quelques locuteurs français. |

*Tableau 1.4 : Quatre niveaux d'erreurs linguistiques produits par les locuteurs non natifs lors de l'acquisition de L2 [Flege, 1995; O'Grady, 2000]*

### 1.4.2. Disparité entre la parole non native et les modèles entraînés

[Tan, 2008] a groupé les niveaux d'erreur d'acquisition L2 selon les trois modèles du système de reconnaissance automatique de la parole. Par exemple, une erreur de phonologie peut être traitée en adaptant les modèles acoustiques. Le tableau 1.5 résume les problèmes de disparité entre la parole non native et les modèles entraînés.

| Caractéristique de la parole non native | Modules de RAP affectés |
|---|---|
| Production approximative des sons de la langue | Modèle acoustique |
| Interférence de la langue maternelle | |
| Fautes de prononciation | Dictionnaire de prononciation |
| Mauvais usage des mots | Modèle de langage |
| Fautes de grammaire | |

*Tableau 1.5 : Disparité entre les erreurs de la parole non native et les modèles en RAP (d'après [Tan, 2008])*

Nous nous intéressons, dans notre contexte de recherche, aux approches d'adaptation des modèles acoustiques pour la parole native et non native. Nous présentons dans la section suivante les différentes techniques d'adaptation autour des modèles acoustiques multilingues pour traiter la parole non native.

### 1.4.3. Différentes techniques d'adaptation pour la parole non native

En raison de la grande variété des accent potentiels pour la parole non native, les études actuelles sont fondées sur les techniques d'adaptation des modèles natifs en utilisant très peu de données de parole non native (au lieu de ré-entraîner les modèles non natifs, ce qui nécessite beaucoup de ressources pour le processus d'entraînement).

Nous présentons dans cette section les différentes techniques d'adaptation des modèles acoustiques multilingues proposées par [Tan, 2007; Wang, 2003] pour traiter la parole non native.

#### *1.4.3.1. MLLR et MAP*

Les techniques d'adaptation des modèles acoustiques *Maximum Likelihood Linear Regression* (MLLR) [Leggetter, 1995] et *Maximum a Posteriori* (MAP) [Gauvain, 1994] sont largement utilisées dans plusieurs domaines de la reconnaissance (du locuteur, de la parole ou de la langue).

L'adaptation MLLR consiste à trouver une matrice de transformation linéaire permettant de convertir le vecteur moyen μ de la gaussienne *m* de l'état *i* d'un HMM en un vecteur moyen adapté μ'. La transformation réalise une combinaison linéaire des gaussiennes du modèle initial pour calculer les nouvelles gaussiennes. La transformation est calculée de manière à maximiser la vraisemblance des données d'adaptation par rapport au modèle, selon un schéma EM (*Expectation/Maximization*). Pour augmenter la robustesse de la méthode, les gaussiennes sont regroupées en classes de régression selon un critère de distance dans l'espace des caractéristiques. Les gaussiennes d'une même classe de régression subissent la même transformation.

Le principe de l'adaptation MAP consiste, quant à lui, à converger de manière itérative vers un ensemble de paramètres optimal (les moyennes des gaussiennes dans le cas présent) en réalisant une combinaison linéaire des paramètres précédents (initialisés avec les paramètres du modèle avant adaptation) avec ceux estimés sur les données d'adaptation, selon le critère du maximum de vraisemblance, selon la formule suivante :

$$\mu_{ig}^{''} = \frac{\sum_{t=1}^{T} \gamma_{ig}(t) o_t}{\tau + \sum_{t=1}^{T} \gamma_{ig}(t)} \bar{\mu}_{ig} + \frac{\tau}{\tau + \sum_{t=1}^{T} \gamma_{ig}(t)} \mu_{ig} \qquad (1.7)$$

$\mu''$ est la moyenne mise à jour ; $\mu$ est la moyenne à l'itération précédente ; $\bar{\mu}$ est la moyenne estimée sur les données d'adaptation ; $o_t$ est le vecteur au temps t d'une séquence issue d'une séquence d'observations de longueur *T* des données d'adaptation ; $\gamma(t)$ est la probabilité de la gaussienne *g* de l'état *i* au temps *t* ; $\tau$ est une constante qui contrôle le poids relatif du modèle initial par rapport aux données d'adaptation.

Avec l'adaptation MAP, contrairement à MLLR, chaque moyenne est mise à jour séparément. Ce fait provoque une différence de comportement entre les deux méthodes :

- MLLR peut être utilisée comme processus d'adaptation de modèles acoustiques, même lorsque la quantité de données d'adaptation est très petite ;
- MAP exige une plus grande quantité de données d'adaptation que MLLR pour être efficace, mais si la quantité de données d'adaptation utilisée est suffisante, l'algorithme MAP sera plus performant que MLLR.

#### *1.4.3.2. Interpolation des modèles acoustiques natifs avec des modèles non natifs*

L'interpolation de modèles acoustiques consiste à effectuer une moyenne pondérée des fonctions de densité multi-gaussienne (*gaussian mixture*) de plusieurs modèles pour produire un seul modèle en sortie. Dans ce cas, nous combinons les modèles acoustiques de la parole

native avec ceux de la parole non native. Les modèles acoustiques natifs sont entraînés sur une grande quantité de données tandis que les modèles non natifs sont entraînés sur très peu de données non natives.

Dans ce cas, il y a deux différents modèles à interpoler, et le modèle acoustique interpolé est défini selon l'équation suivante :

$$MA_{\text{inter}}(O) = w_{native}.MA_{native}(O) + w_{non-native}.MA_{non-native}(O) \quad (1.8)$$

où $w_{native} + w_{non-native} = 1$, $MA(O)$ est un modèle acoustique, $O$ est un vecteur d'observation des caractéristiques acoustiques, et $w$ est le poids d'interpolation.

Selon les études de [Wang, 2003], la technique d'interpolation ci-dessus est notablement plus efficace que le re-entraînement des modèles acoustiques mentionné dans [Livescu, 1999] et [Uebler, 1999] (section 1.3.3).

#### *1.4.3.3. Polyphone decision tree specialization (PDTS)*

Pour la modélisation acoustique dépendante du contexte, il y a de grandes différences entre la parole de locuteurs natifs et celle des locuteurs non natifs. En effet, quand nous décodons la parole non native avec un modèle acoustique dépendant du contexte (figure 1.5) qui a été construit à partir de la parole native, la performance est dégradée très significativement, parce que l'arbre de décision ne représente pas, de façon très précise, le contexte de la parole non native.

La meilleure solution est de construire l'arbre de décision pour la parole non native à partir de zéro. Le problème ici est qu'il nous faut suffisamment de données non natives pour construire cet arbre.

Dans le cas où les données non natives sont insuffisantes pour la construction de l'arbre « non natif », [Wang, 2003] propose d'adapter une approche nommée *Polyphone Decision Tree Specialization* [Schultz, 2000], créée au départ pour résoudre le problème de la construction d'un arbre de décision multilingue. Pour la parole non native, on sélectionne la meilleure correspondance acoustique pour chaque mot lors de l'alignement, ensuite on génère une liste de nouveaux polyphones. Les nouveaux polyphones sont ensuite intégrés dans l'arbre de décision, avec des branches élaguées à la dernière feuille de données origines (voir la figure 1.9). Nous obtenons, à la fin, un arbre de décision « non natif ».

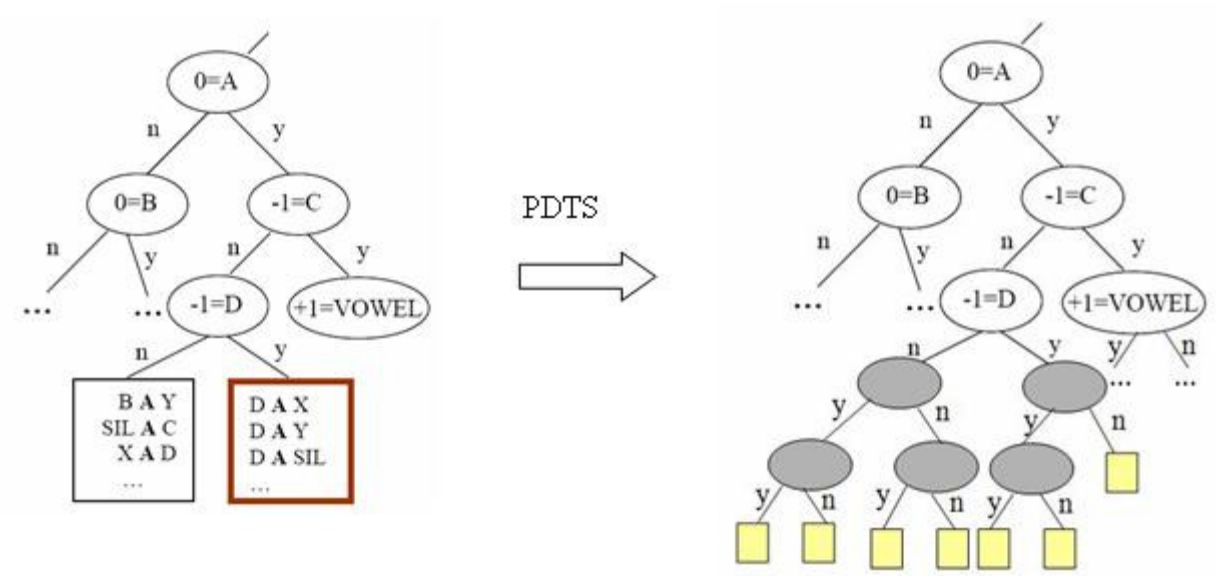



*Figure 1.9 : Utilisation de la technique PDTS pour transformer l'arbre de décision de la parole native en un arbre de décision pour la parole non native [Wang, 2003]*

### *1.4.3.4. Approche hybride: interpolation et fusion*

Dans la section 1.4.1.3, il est dit que les locuteurs empruntent des caractéristiques linguistiques (acoustique, phonétique, vocabulaire, grammaire, etc.) de leur langue maternelle (L1) dans la parole non native. Une bonne idée peut donc être de considérer les modèles acoustiques de L1 et les modèles acoustiques de L2 pour créer les modèles acoustiques non natifs de la langue L2 parlée par les locuteurs de langue maternelle L1. Cela a été mis en évidence par [Tan, 2007] qui a proposé une approche hybride d'adaptation des modèles acoustiques pour la parole non native ne nécessitant pas de données d'adaptation, et où seuls les modèles acoustiques de L1 et L2 sont utilisés. Cette approche hybride est appelée « interpolation et fusion ».

Le processus d'adaptation hybride proposé consiste à interpoler et fusionner les gaussiennes cibles (les gaussiennes dans le modèle acoustique L2) et sources (les gaussiennes dans le modèle acoustique L1).

Lorsque la distance euclidienne (*dist( )*) entre une gaussienne cible ($g_{cible,sn}$) et les gaussiennes sources associées ($g_{source,sn}$) est inférieure à un seuil (*dist*), alors leurs vecteurs moyens (means), la matrice de covariance (*variance*s) et la distribution multi-gaussienne ($\omega$, *mixture weights*) seront interpolées (équation 1.9). Sinon, pour les gaussiennes sources qui sont loin des autres gaussiennes cibles (équation 1.10) ou pour les gaussiennes cibles sans aucune gaussienne source associée (équation 1.11), on conserve les deux distributions

(fusion). En cas de fusion, la distribution multigaussienne sera ajustée selon le poids d'interpolation (*w*). Le seuil de distance (*dist*) peut être calculé, par exemple, en mesurant la distance moyenne entre les gaussiennes, puis en le multipliant par une constante. [Tan, 2007] formule l'approche hybride (interpolation et fusion) de deux modèles acoustiques (L1 et L2) comme suit :

$$g_{inter,sn} = (1-w).g_{cible,sn} + w.g_{source,sn}\ ,\ g_{source,sn} \neq \varnothing,$$
$$d(g_{cible,sn}, g_{source,sn}) \leq dist \tag{1.9}$$

$$g_{inter,sn} = g_{source,sn}\ ,\ \omega_{inter,sn} = w.\omega_{source,sn}\ ,\ g_{cible,sn} \neq \varnothing,$$
$$d(g_{cible,sn}, g_{source,sn}) > dist \tag{1.10}$$

$$g_{inter,sn} = g_{cible,sn}\ ,\ \omega_{inter,sn} = (1-w).\omega_{cible,sn}\ ,\ g_{source,sn} = \varnothing \tag{1.11}$$

La figure 1.10 donne un exemple d'application d'une adaptation hybride pour créer le modèle acoustique de la parole non native française parlée par les locuteurs vietnamiens. Dans l'espace acoustique (figure 1.10), la première gaussienne de l'état 1 du phonème français /p/ ($p_{FR,e1g1}$) est interpolée avec les deux premières gaussiennes de l'état 1 du phonème vietnamien /p/ ($p_{VN,e1g1}$ et $p_{VN,e1g2}$). Au contraire, les deux gaussiennes $p_{FR,e1g2}$ et $p_{VN,e1g3}$ qui sont loin l'une de l'autre sont conservées (fusion) (les deux gaussiennes sont conservées mais leurs distributions multi-gaussiennes sont recalculées).

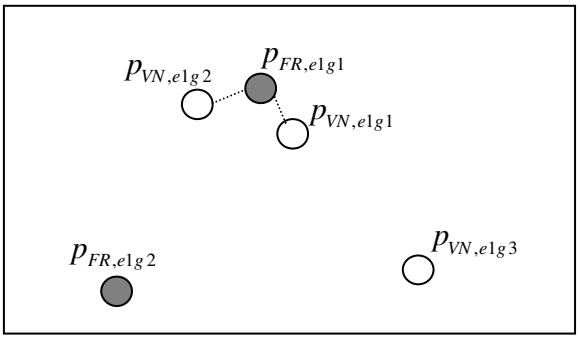



*Figure 1.10 : Espace acoustique - un exemple d'interpolation et fusion de phonèmes de gaussiennes pour les phonèmes /p/ français et vietnamien [Tan, 2007]*

Cette dernière approche, qui fait une adaptation des modèles acoustiques sans utiliser de données d'adaptation, est bien adaptée à notre contexte d'étude mentionné au début du chapitre. Pour cette raison, nous utiliserons cette adaptation hybride (interpolation et fusion) des modèles acoustiques, mais de manière non supervisée (sans connaître a priori la langue

L2 et L1 du locuteur). Cela constitue une des contributions de ce travail de thèse et sera détaillé dans les chapitres 3 et 4 de ce mémoire.

## 1.5. Conclusion

Dans ce chapitre, nous avons présenté le contexte et la motivation de notre thèse. Nous avons par la suite abordé le principe de la reconnaissance de la parole par l'approche statistique, et décrit les différentes composantes d'un système de RAP standard : modèles acoustiques, modèles de langage, dictionnaire de prononciation. Enfin, nous avons présenté très brièvement le principe des modèles acoustiques multilingues et l'utilisation qui peut en être faite dans différents domaines du traitement de la parole. En fin de chapitre, nous avons présenté les caractéristiques de la parole native et non native et les différentes techniques d'adaptation des modèles acoustiques pour traiter la parole non native.

# Chapitre 2
# Acquisition & analyse d'un corpus de type « réunions multilingues »

## 2.1. Introduction

Pour tester les approches proposées, il est nécessaire d'avoir des données de test contenant de la parole native et non native parlée par des locuteurs venant de différentes origines. Pour atteindre cet objectif, nous avons construit un corpus qui contient de la parole issue de réunions multilingues. Dans une réunion multilingue interviennent des locuteurs venant de différents pays et qui sont de langues maternelles différentes. Dans ce type de réunion, les locuteurs peuvent parler soit leur langue maternelle (parole native) soit des langues étrangères (parole non native). Ce chapitre présente, étape par étape, le processus d'acquérir du corpus de réunion multilingue dont nos données de test sont ensuite extraites. Ce corpus est appelé « *MICA Multilingual Meeting Speech Corpus* » (MICA-MultiMeet) car il a été recueilli dans la salle de réunion du centre de recherche *MICA*[7]. MICA-MultiMeet contient de la parole native et non native pour quatre langues : l'anglais (EN), le français (FR), le khmer[8] (KH) et le vietnamien (VN). À la fin du chapitre, nous analysons certaines confusions de phonèmes des locuteurs non natifs à travers les langues en question (EN, FR et VN).

## 2.2. Acquisition du corpus « MICA-MultiMeet »

En vue du développement d'un système de RAP, le corpus de parole est divisé typiquement en trois groupes de données : les données d'entraînement, d'adaptation et de test. Cela est généralement vrai pour un corpus de parole native. Mais, pour développer un système de reconnaissance automatique de parole non native, il est difficile de collecter suffisamment de données pour entraîner des modèles « non natifs ». En outre, il existe un trop grand nombre de groupes (origines) de locuteurs non natifs qui peuvent être impliqués dans une conversation (par exemple la langue anglaise peut être parlée par des locuteurs français,

[7] Multimédia, Informations, Communication et Applications (http://www.mica.edu.vn)
[8] Khmer : la langue officielle du Cambodge

vietnamiens, cambodgiens, chinois, etc.). Ainsi, notre corpus « MICA-MultiMeet » n'est développé que pour tester les approches proposées et/ou adapter les modèles existants. Puisque la transcription est une tâche qui prend énormément de temps et d'efforts pour collecter une base de données de parole, le corpus « MICA-MultiMeet » a été créé à partir de scénarios textuels prédéfinis. Cela signifie que, lors de la réunion, les participants utilisent un texte (script) de conversation tout en essayant de maintenir un rythme de parole aussi normal que lors d'une vraie réunion.

La figure 2.1 illustre les trois étapes du processus d'acquisition du corpus « MICA-MultiMeet »:

- pré-enregistrement (préparation des scénarios textuels, diagnostic de la qualité du signal) ;
- enregistrement (outils d'enregistrement et participants) ;
- transcription (synchronisation, vérification et extraction des textes et signaux).

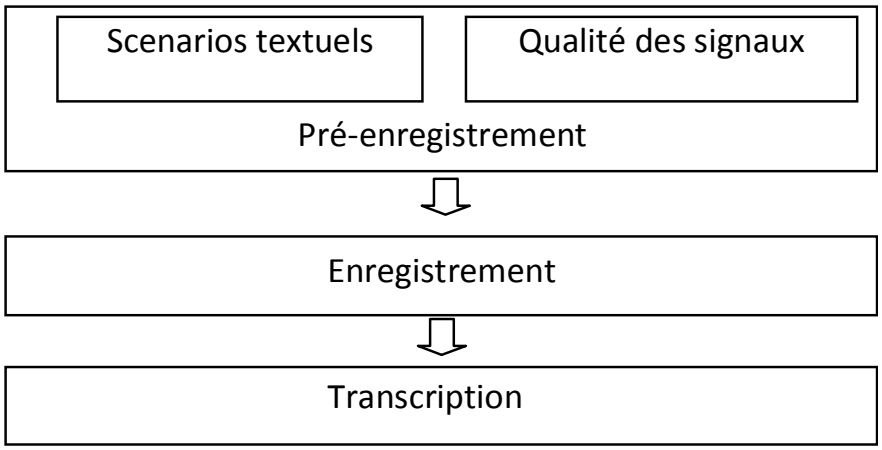



*Figure 2.1 : Processus d'acquisition du corpus « MICA-MutiMeet »*

### 2.2.1. Scénarios textuels

Avant l'enregistrement, nous avons préparé 19 scénarios textuels de réunion multilingue. Le contexte d'un scénario est une discussion ou une présentation d'un sujet général ou scientifique. Dans chaque scénario, un locuteur continue généralement à parler une langue à moins qu'un autre ne se mette à parler dans une nouvelle langue dans le dialogue. Dans ce dernier cas, un changement de langue se produit quand les locuteurs actifs dans le dialogue ont changé.

**Eric (FRfr):** Bonjour! Aujourd'hui, Vattana va nous présenter un sujet intitulé « gestion des documents multimédia pour les cours en ligne ». Bon! Je lui laisse la parole.

**Sethserey (KHkh):** អេរិចចង់អោយឯងនិយាយពីប្រធានបទឥឡូវនេះ (*Eric veut que tu parles de ton sujet maintenant*)

**Vattana (ENkh)**: Ok! First of all, I would like to present the objective of my thesis... (*Ok ! Tout d'abord, je voudrais présenter l'objectif de ma thèse…*)

**Linh (VNvn)** : Vattana ! tiếng Việt de bằng de tôi de cho de thích de giải de thể de có de bạn ? (*Vattana ! pouvez-vous me l'expliquer en vietnamien ?*)

*Tableau 2.1 : Exemple d'un scénario textuel de réunion multilingue*

Notez que, dans les parenthèses au début des phrases du tableau 2.1, les étiquettes en majuscules dénotent les langues parlées (langues cibles) et les étiquettes en minuscules dénotent les langues maternelles des locuteurs (langues sources). Par exemple, ENkh indique de la parole non native anglaise parlée par un locuteur de la langue source « khmer ». D'autre part, les textes français en italique dans les parenthèses à la fin de chaque phrase sont simplement la traduction correspondante de la langue parlée (ils ne sont pas lus par les locuteurs pendant l'enregistrement).

Pour manipuler pleinement les scripts complexes du vietnamien et du khmer, tous les scénarios textuels sont encodés en *Unicode* (format *UTF-8*) [Aliprand, 2003]. D'ailleurs, pour avoir la représentation en mots et en phonèmes dans le processus de transcription, nous segmentons les phrases du khmer en mots (normalement, un texte khmer est écrit sans espaces entre les mots). Par exemple, dans le tableau 2.1, la phrase អេរិចចង់អោយឯងនិយាយពីប្រធានឥឡូវនេះ devrait être segmentée comme អេរិច ចង់ អោយ ឯង និយាយ ពី ប្រធានបទ ឥឡូវនេះ (Éric veut que tu parle du sujet maintenant). Pour segmenter le texte khmer en mots, nous avons utilisé un outil de segmentation automatique de texte en mots développé au sein de l'équipe GETALP du laboratoire d'informatique de Grenoble (LIG) [Seng, 2008b].

Enfin, nous demandons aux personnes de vérifier leur langue maternelle écrite dans les scénarios textuels avant de les donner aux personnes impliquées (locuteurs) dans la réunion. Pour que les locuteurs soient bien préparés à leurs rôles, le scénario textuel leur est envoyé quelques jours avant l'enregistrement.

### 2.2.2. Diagnostic concernant la qualité du signal

Avant de procéder à l'acquisition du corpus, il est indispensable d'étudier la qualité du signal de parole enregistré dans l'environnement de la salle de réunion pour savoir si cette qualité sera suffisante pour nos futures expérimentations.

Le rapport signal sur bruit (RSB)[9] est un élément important dans la détermination de la qualité des données audio. Cela est particulièrement important en reconnaissance vocale, car il est bien connu que la performance de reconnaissance est fortement influencée par le RSB (la plus grande valeur de RSB correspond à la meilleure performance). En décibel, ce rapport est défini comme le rapport entre l'amplitude du signal ($A_{signal}$) et l'amplitude du bruit ($A_{bruit}$) corrompant ce signal (voir équation 2.1). Généralement, un RSB plus grand que 40 dB est adéquat pour des études phonétiques.

$$RSB = 20 * log_{10} \frac{A_{signal}}{A_{bruit}} \tag{2.1}$$

La pente spectrale ($P$) de la voyelle [Rosen, 1991] est aussi une méthode utilisée pour évaluer la qualité de la parole. En théorie, $P$ doit être entre -12 dB/octave et -6 dB/octave. Elle est formulée comme suit :

$$P = (I1 - I2) / log_2(F2/F1) \tag{2.2}$$

où $F1$ et $F2$ sont les premier et deuxième formants d'une voyelle, et ont l'intensité $I1$ et $I2$ respectivement.

Dans notre diagnostic du signal, nous avons extrait 30 voyelles /a/ et /i/ *des signaux* enregistrés dans la salle de réunion du centre de recherche MICA. L'outil « *Praat* » [Boersma, 2001] a été utilisé pour mesurer le rapport RSB et la pente spectrale. Nous observons finalement que notre qualité de signal de parole est suffisante (60 dB <= RSB <= 65 dB ; -10 dB/octave <= $P$ <= -6.5 dB).

En outre, selon Huang [Huang, 2001], une fréquence d'échantillonnage de 16 kHz est suffisante pour représenter intelligiblement la parole humaine et il démontre qu'une fréquence d'échantillonnage supérieure à 16 kHz n'améliore pas la performance d'un système de reconnaissance vocale. Ainsi, nous avons fixé l'échantillonnage des signaux à enregistrer à 16 kHz et ils sont encodés en WAV mono canal avec une quantification sur 16 bits.

[9] http://en.wikipedia.org/wiki/Signal-to-noise_ratio

### 2.2.3. Enregistrement

#### *2.2.3.1. Matériels*

Tous les enregistrements ont été faits dans la salle de réunion du centre de recherche MICA. La salle est équipée par un système de communication multimédia de haute performance pour assurer le travail de collaboration à distance (réunions, vidéoconférences, etc.). Ce système de communication a été créé dans le cadre du projet SIAM (Salle Intelligente pour les Applications en Multimédia). Toutes les références techniques sont détaillées dans [Lazzarotto, 2007].

L'outil d'enregistrement se compose de :

- quatre micros-cravate (modèle « *Sennheiser EW300G2 »*) ;
- un outil d'acquisition multifonctionnelle des signaux (*multifunctional data acquisition, DAQ*) de marque « NI BC-2110 » qui sert à relier les quatre micros-cravate à la carte son de l'ordinateur ;
- une carte son de haute qualité « *soundMAX Integrated Digital HD audio »* ;
- un logiciel d'enregistrement (appelé « *SmartRoom* ») développé au centre de recherche MICA ; ce logiciel est utilisé pour assurer l'enregistrement parallèle des quatre micros-cravate.

#### *2.2.3.2. Participants*

Lors de chaque réunion, nous avons trois ou quatre locuteurs et un assistant d'enregistrement. Les locuteurs savent parler au moins une langue étrangère. Chaque locuteur porte un micro-cravate et il parle à partir du scénario textuel qui lui a été fourni. Les locuteurs sont assis autour d'une grande table ronde à une distance de 1,5 mètre environ les uns des autres. L'assistant note aussi les cas où il y a des bruits produits pendant l'enregistrement. À chaque enregistrement, un scénario est répété cinq fois ; les quatre derniers seulement seront conservés. La première occurrence permet de vérifier l'état des matériels et des logiciels d'enregistrement, et de régler le système d'enregistrement dans le cas où l'intensité de la parole des locuteurs est trop forte ou trop faible. La durée totale d'une réunion peut varier entre 15 minutes et 30 minutes.

### 2.2.4. Transcription

L'outil *Transcriber* [Barras, 2001] a été utilisé dans le processus de transcription. La transcription a été validée selon une approche en trois étapes comme illustré dans la figure 2.2.

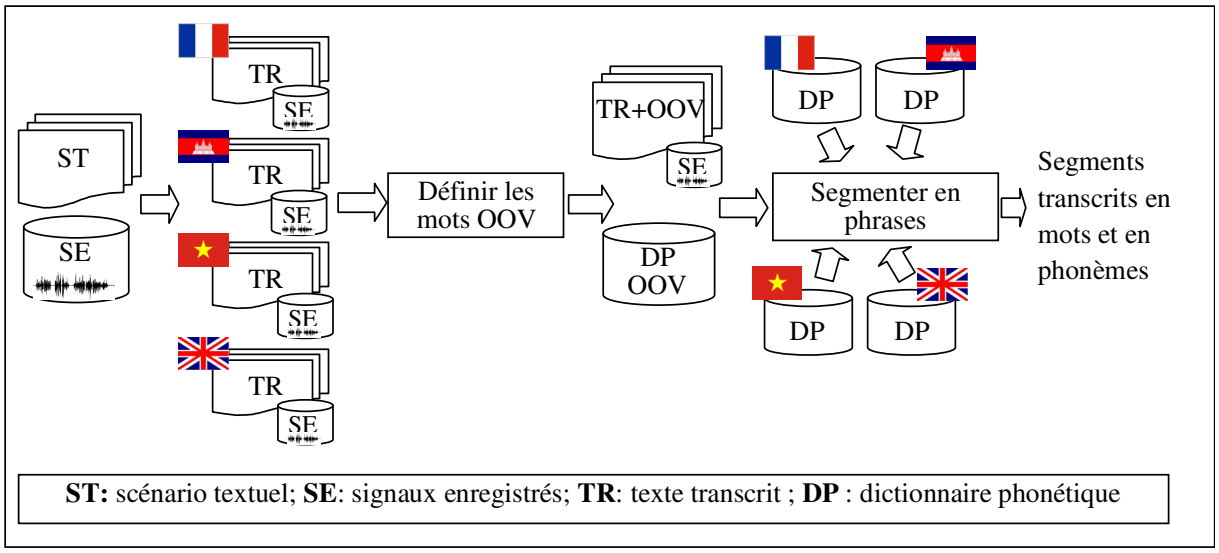



*Figure 2.2 : Processus de transcription des signaux du corpus « MICA-MutiMeet »*

Premièrement, comme il y avait des erreurs faites par les locuteurs dans leurs discours (prononciation incorrecte, répétition, insertion des bruits vocaux, etc.), les scénarios des textes et les signaux d'enregistrement ont dû être vérifiés, modifiés et validés par les transcripteurs natifs des langues parlées. Pour atteindre cet objectif, les scénarios textuels ont été synchronisés et segmentés langue par langue. Ensuite, tous les segments d'une langue parlée ont été donnés à son transcripteur natif. Pour assurer la synchronisation des textes et des signaux, les transcripteurs doivent assurer la cohérence entre les textes et les signaux qu'ils ont entendus. D'ailleurs, pour aider le phonéticien à trouver facilement les mots exceptionnels, les transcripteurs l'aident à marquer les mots hors vocabulaire (*OOV*) comme les noms propres, les mots mal prononcés, les bruits vocaux (toux, souffle, etc.) et les bruits extérieurs (par exemple le bruit de la porte).

Deuxièmement, après la synchronisation et la vérification des données textuelles et de la parole par les transcripteurs natifs, le phonéticien crée un vocabulaire phonétique d'*OOV* (mot + sa représentation phonétique) en détectant premièrement les marques d'*OOV* fournies par les transcripteurs. Il est important de souligner que, pour assurer une représentation phonétique unique, un seul phonéticien a été sollicité pour faire l'annotation et la transcription de tous les mots d'*OOV*. De plus, nous utilisons le format de X-SAMPA proposé dans [Wells, 1995]

comme clef de représentation phonétique pour tous les textes transcrits et pour tous les dictionnaires phonétiques.

Troisièmement, l'extraction des segments, selon les tours de parole, et l'extraction des textes transcrits avec les signaux correspondants, est faite automatiquement. Les textes extraits sont non seulement disponibles dans le format orthographique original (séquence des mots), mais également dans le format phonétique X-SAMPA (nous convertissons les mots en phonèmes en employant le dictionnaire phonétique X-SAMPA du français, de l'anglais, du vietnamien, du khmer et notre liste de mots *OOV*).

## 2.3. État actuel du corpus « MICA-MultiMeet »

La quantité totale de parole enregistrée est d'environ six heures (cinq heures de signal transcrit et une heure de signal non transcrit) parlées par 18 locuteurs (14 sujets masculins et quatre sujets féminins). Ces signaux sont répartis en parole native et non native pour quatre langues : l'anglais (EN), le français (FR), le khmer (KH) et le vietnamien (VN). La taille moyenne de chaque segment (tour de parole) de signal transcrit est d'environ 3,5 secondes. Le nombre moyen de mots parlés dans un segment varie en fonction des langues.

| | **# locuteurs (Sexe)** | **Lang-KH** | **Lang-VN** | **Lang-FR** | **Lang-EN** | **Total par origine** |
|---|---|---|---|---|---|---|
| **Locuteur-KH** | 3(M) | 570 | 452 | 1822 | 3452 | 6296 |
| **Locuteur-VN** | 2(F), 3(M) | 0 | 1747 | 1177 | 675 | 3599 |
| **Locuteur-FR** | 1(F), 4(M) | 0 | 1550 | 2797 | 1370 | 5717 |
| **Locuteur-EN** | 1(M) | 0 | 590 | 584 | 911 | 2085 |
| **Total par langue** | 14 | 570 | 4339 | 6380 | 6408 | **17697** |

*Tableau 2.2 : Durée totale des signaux transcrits (valeurs en secondes)*

Le tableau 2.2 présente la quantité totale de signal transcrit disponible dans le corpus « MICA-MultiMeet ».

Nous observons que 66 % du signal transcrit du corpus « MICA-MultiMeet » sont de la parole non native. En outre, dans notre campagne d'enregistrement du corpus, nous avons rencontré une certaine difficulté pour trouver des locuteurs des langues française et anglaise qui peuvent parler les langues peu ou très peu dotées (les langues qui disposent de peu de ressources numériques, voire [Berment, 2004]) comme la langue vietnamienne et la langue khmère. En conséquence, la parole non native en langue khmère n'est pas disponible dans l'état actuel du corpus « MICA-MultiMeet ».

## 2.4. Corpus de test

Pour étudier et évaluer les impacts de la langue maternelle des locuteurs sur la parole non native (chapitre 3) et l'adaptation autonome des modèles acoustiques que nous avons proposée (chapitre 4), nous avons extraits des données de test (du signal transcrit) à partir du corpus « MICA-MultiMeet » mentionné dans la section précédente. Dans le cadre de notre thèse, nous avons décidé de ne pas considérer les segments qui contiennent plusieurs langues parlées (*code-mixing*). Nous nous intéressons seulement aux segments homogènes de parole native ou non native. C'est pourquoi, dans notre corpus de test, chaque segment de signal transcrit ne contient qu'une langue parlée (parole native ou non native).

En raison de l'absence de locuteurs non natifs en langue khmère (voir le tableau 2.2), le corpus de test, pour nos études, contient uniquement de la parole native et non native de l'anglais, du français et du vietnamien.

Enfin, nous ne sélectionnons que les segments de parole native et non native des langues EN, FR et VN d'une durée supérieure à cinq secondes pour nos expériences dans les chapitres suivants. En conséquence, le corpus de test contient environ 52 minutes de parole transcrite (voir tableau 2.3), dont la parole non native représente 69 % [Sam, 2010c].

| **Langue** | **Parole native/non native** | **Total par langue** |
|---|---|---|
| EN | ENen=239 ; ENfr=715 ; ENvn=241 | 1195 |
| FR | FRfr=241 ; FRen=253 ; FRvn=486 | 980 |
| VN | VNvn=235 ; VNen=215 ; VNfr=483 | 933 |
| **Total** | | **3108** |

*Tableau 2.3: Quantité de données de test (valeur en secondes)*

## 2.5. Analyse des confusions de phonèmes des locuteurs non natifs à travers les langues (*cross-language transfer*)

Avec le corpus de réunion multilingue (MICA-MultiMeet) qui contient environ 70 % de parole non native, une analyse phonétique interlingue des locuteurs non natifs est nécessaire. L'objectif de l'analyse est non seulement de savoir comment les locuteurs non natifs réagissent aux phonèmes de la langue cible qui n'existent pas dans la langue source (langue

maternelle), mais peut être aussi de nous permettre de connaître la relation entre la langue maternelle (langue source, L1) des locuteurs et la langue parlée (langue cible, L2).

Comme il n'y a pas de locuteur non natif en langue khmère dans le corpus « MICA-MultiMeet », l'analyse phonétique interlingue des locuteurs non natifs ne sera effectuée que pour les langues anglaise, française et vietnamienne.

Avant d'effectuer cette analyse, nous présentons tout d'abord les phonèmes communs et les phonèmes particuliers des trois langues impliquées, à l'aide de leurs représentations phonétiques dans le tableau de l'Alphabet Phonétique International (API) [IPA, 1999].

### 2.5.1. Alphabet Phonétique International (API)

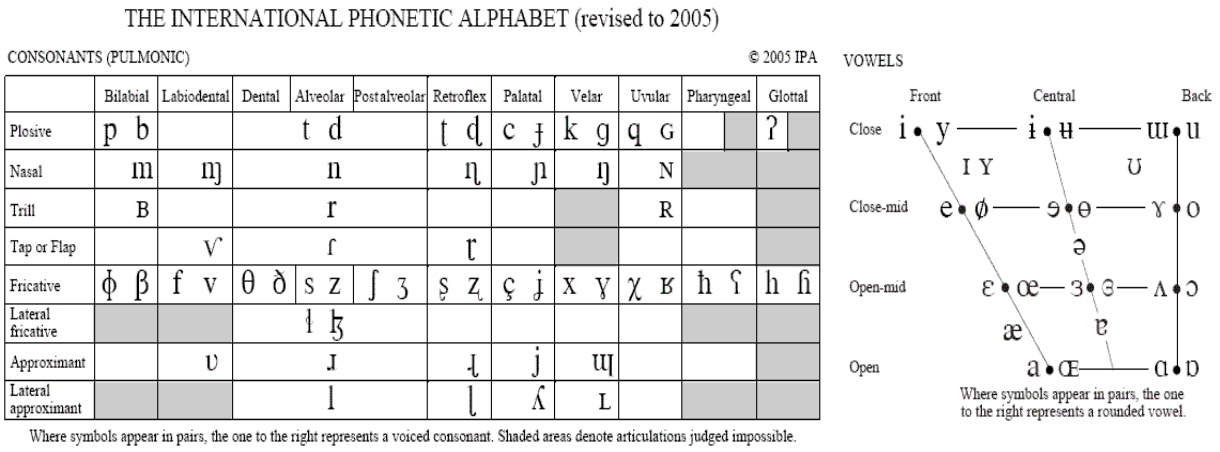



*Figure 2.3 : API pour les consonnes et les voyelles*

L'Alphabet Phonétique International (API) est une des représentations phonétiques les plus utilisées. L'API a été créé par des groupes de phonéticiens pour des études linguistiques. L'API recense toutes les unités phonétiques (phonèmes) de la parole selon le système vocal humain. Les consonnes sont classées selon la position d'articulation, le type d'articulation, le voisement, et les voyelles sont classées selon la position de la langue et l'arrondissement des lèvres. Le lecteur trouvera une illustration du tableau d'API plus détaillée dans l'annexe A.

Le tableau 2.4 présente tous les phonèmes des langues française, anglaise et vietnamienne, classés selon qu'ils sont particuliers ou commun à chaque langue. Il est important de préciser que les listes de phonèmes présentées dans le tableau 2.4 sont issues d'études précédentes sur les modèles acoustiques de la parole lue française (BREF [Lamel, 1991]), vietnamienne (VNSpeechCorpus [Le, 2004]) et anglaise (WSJ [Paul, 1992]). Pour être lisible par les lecteurs, surtout les phonéticiens, nous présentons les phonèmes au format API au lieu de X-SAMPA, bien que le format X-SAMPA soit utilisé pour nos expérimentations.

Dans le tableau 2.4, nous observons que les phonèmes communs au couple de langues anglaise-français sont plus nombreux que ceux du couple de langues anglais-vietnamien et que ceux du couple de langues français-vietnamien. Par exemple, il y a 28 phonèmes communs entre l'anglais et le français, mais il n'y a que 23 phonèmes communs entre l'anglais et le vietnamien et que 24 phonèmes communs entre le français et le vietnamien. Cette variation est peut être due à une plus grande distance entre les langues (les langues austro-asiatiques sont très éloignées des langues indo-européennes)[10].

| **Groupe** | **Phonèmes (API)** | **FR** | **EN** | **VN** | **∑ par langue** | **∑ par groupe** |
|---|---|---|---|---|---|---|
| Particulier | ɑ, ɥ, ʁ, ɑ̃, ɛː, ɛ̃, ø, ɔ̃, œ̃, œ, oɔ, øœ, y | X | | | 13 | 40 |
| | ʌ, ɪ, ʊ, aɪ, æ, ð, dʒ, ɔɪ, r, tʃ, θ | | X | | 11 | |
| | ɤ, ɣ, ɯ, ʂ, zᵲ, ɯɤ, aχ, c, cʰ, ɛχ, ie, ɔχ, tʰ, uo, χ, ɤχ | | | X | 16 | |
| Commun | ʃ, aː, g, ʒ, ə | X | X | | 5 | 29 |
| | ɲ | X | | X | 1 | |
| | a, b, d, e, ɛ, f, h, i, j, k, l, m, n, ŋ, o, ɔ, p, s, t, u, v, w, z | X | X | X | 23 | |
| *Monolingue ∑ = 121* | | 42 | 39 | 40 | | |

*Tableau 2.4 : Ensemble de phonèmes particuliers et communs des langues impliquées (selon API)*

### 2.5.2. Analyse phonétique interlingue selon des modèles statistiques

L'analyse phonétique interlingue du locuteur non natif peut se faire soit par une approche à base de connaissances (*knowledge-based*) soit par une approche automatique à base de modèles statistiques (*data-driven*). La première nécessite beaucoup d'effort, de temps et de ressources. En outre, elle exige une personne avec une connaissance linguistique spécialisée dans le domaine pour analyser la parole non native.

Par ailleurs, les phonéticiens impliqués doivent posséder la connaissance des langues visées. Quelques phonèmes, par exemple le phonème /ʁ/ français, peuvent également être difficiles à analyser pour un phonéticien qui n'est pas familier avec la langue française.

L'approche automatique à base de modèles statistiques présente les avantages d'être rapide, et elle peut être assez précise si les modèles utilisés ont été créés avec suffisamment de données.

Dans notre travail, nous utilisons une approche automatique à base de modèles statistiques appelée « matrice de confusion de phonèmes ». L'idée est de trouver la confusion ou

[10] http://www.freelang.com/familles/index.php#top

l'asymétrie entre les phonèmes prononcés par les locuteurs non natifs et les vrais phonèmes de la langue parlée (langue cible ou langue seconde L2).

### *2.5.2.1. Le processus de la matrice de confusion de phonèmes*

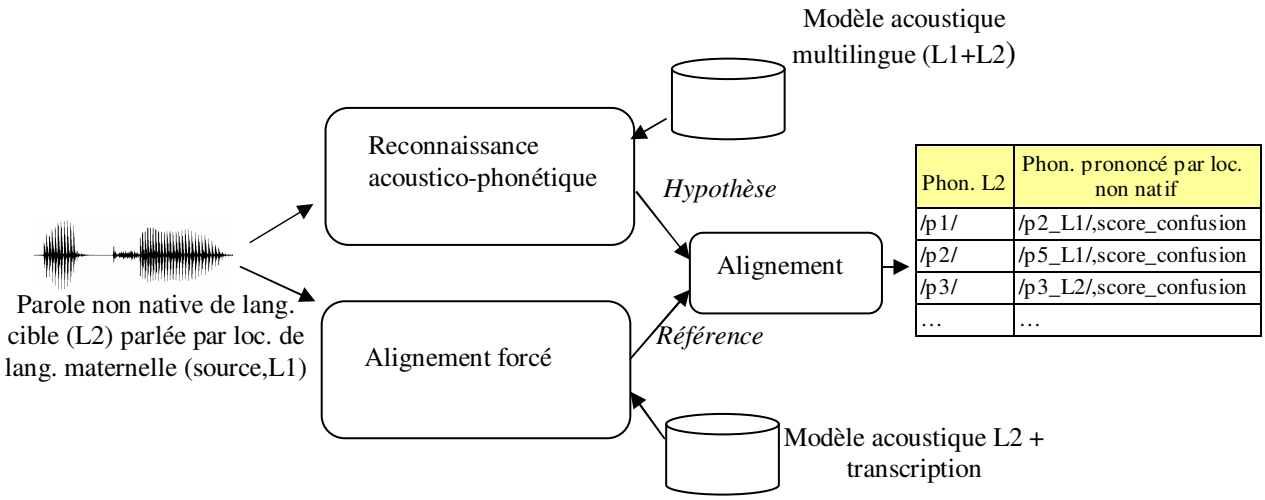



*Figure 2.4 : Processus pour trouver les phonèmes prononcés par les locuteurs non natifs correspondant aux vrais phonèmes de la langue parlée à base de la matrice de confusion de phonème*

Le processus de la matrice de confusion de phonème est réalisé en utilisant trois modules.

**Reconnaissance acoustico-phonétique (hypothèse) :**

Les segments de parole non native sont tout d'abord décodés par un système de reconnaissance acoustico-phonétique. Comme les locuteurs non natifs sont influencés par leur langue maternelle quand ils apprennent une nouvelle langue (selon [Flege, 1997]), la reconnaissance acoustico-phonétique utilisée doit être capable de reconnaître à la fois les phonèmes de la langue cible et ceux de la langue d'origine du locuteur. Ainsi, nous avons créé trois systèmes acoustico-phonétiques qui utilisent trois modèles acoustiques multilingues (MA-Mult) français-anglais, français-vietnamien et anglais-vietnamien pour décoder respectivement la parole française parlée par les locuteurs anglais (et inversement), la parole française parlée par les locuteurs vietnamiens (et inversement) et la parole anglaise parlée par les locuteurs vietnamiens (et inversement). Les modèles acoustiques indépendants du contexte (CI) des langues anglaise, française et vietnamienne à combiner sont entraînés à partir des corpus WSJ [Paul, 1992], BREF [Lamel, 1991] et VNSpeechCorpus [Le, 2004] respectivement. La combinaison des modèles acoustiques s'est faite en utilisant la méthode des langues séparées *ML-sep* [Schultz, 2001]. Cela signifie qu'il n'existe pas de données à

partager, à travers les trois modèles acoustiques combinés. L'outil *sphinx3* a été utilisé pour décoder toutes les sortes de parole non native. Le détail de ces modèles acoustiques et de la méthode de combinaison des modèles acoustiques sera présenté dans le chapitre 3.

**Alignement forcé (référence) :**

Les séquences phonétiques de référence sont obtenues par alignement forcé (en utilisant l'outil sphinx3_align[11]). Cet outil aligne les séquences phonétiques correspondant à un segment de parole non native en utilisant le modèle acoustique de la langue cible et la transcription phonétique correspondant à ce segment.

**Alignement de la référence avec l'hypothèse :**

Finalement, nous pouvons construire la matrice de confusion entre les phonèmes d'hypothèse (résultats de la reconnaissance acoustico-phonétique multilingue source-cible) et les phonèmes de référence en langue cible (résultats d'alignement forcé). La matrice de confusion d'un segment de signal est, en fait, créée en se basant sur la comparaison entre la suite de phonèmes étiquetés en langue source (hypothèse) avec la suite de phonèmes étiquetés en langue cible (référence) sur un axe temporel pour obtenir la correspondance phonétique trame par trame (figure 2.5). Le lecteur trouvera une explication plus détaillée sur l'approche de matrice de confusion selon l'alignement temporel de phonèmes en langues source/cible dans [Le, 2006]. On obtient ainsi une matrice *A(M, N)* qui est la matrice de confusion de phonèmes, avec $0 <= A_{ij} <= 1$, mesure de confusion entre le phonème $t_j$ en langue cible et le phonème $s_i$ en langue source. La distance entre phonèmes peut alors s'écrire simplement :

$$d(s_i, t_j) = A_{ij} \tag{2.3}$$

où $Aij \in [0, 1]$ et $i \in [1..M]$, $j \in [1..N]$

[11] http:\\cmusphinx.sourceforge.net

| trame | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| référence | sil | | c | i | | h | ɔ | j | | a | | j | sil | | v | | ɤ̆ | | j | |
| hypothèse | sil | | | s | i | ɔ | | i | a | | | i | sil | | v | | | a | | sil |

*Figure 2.5 : Alignement temporel en langue source/cible [Le, 2006]*

### *2.5.2.2. Résultat des confusions de phonèmes des locuteurs non natifs*

Nous avons extrait toutes la parole non native en anglais (ENfr, ENvn), français (FRen, FRvn) et vietnamien (VNen, VNfr) du tableau 2.2 pour obtenir une matrice de confusion.

Dans le tableau 2.5a, la première colonne liste tous les phonèmes de la langue française (FR) alors que la deuxième et la troisième colonne présentent (avec les meilleurs scores de confusion entre parenthèses) les phonèmes prononcés par les locuteurs anglais (FRen) et vietnamiens (FRvn) respectivement. La première colonne de la dernière ligne du tableau présente le nombre total de phonèmes de la langue cible. Les deuxième et troisième colonnes de la dernière ligne du tableau listent le nombre total de phonèmes français que les locuteurs anglais et vietnamiens peuvent prononcer, le nombre total des phonèmes empruntés à la langue d'origine du locuteur (anglais ou vietnamien) et enfin le nombre total de phonèmes de la langue cible qui n'ont jamais été reconnus (une raison possible en est que les données de parole non native sont insuffisantes pour calculer les scores de confusion associés à certains phonèmes plus rares ; une autre raison peut être que ces phonèmes sont très difficiles à prononcer pour les locuteurs non natifs).

La même grille de lecture est valable pour le tableau 2.5b (les phonèmes anglais prononcés par les locuteurs français et vietnamiens) et le tableau 2.5c (les phonèmes vietnamiens prononcés par les locuteurs français et anglais).

a.

| FR | FRen | FRvn |
|---|---|---|
| a | a_FR(0.105) | a_VN(0.266) |
| ɑ | a ː _FR(0.364) | ɑ_FR(0.2) |
| ɑ̃ | ɑ̃_FR(0.411) | ɑ̃_FR(0.125) |
| a ː | ??? | a_VN(1) |
| b | b_FR(0.667) | v_VN(0.224) |
| ɔ | u_FR(0.339) | ɔχ_VN(0.123) |
| ɔ̃ | o_EN(0.248) | o_VN,ɔ_VN (0.095) |
| d | d_FR(0.198) | d_FR(0.096) |
| e | ɛ_FR(0.122) | e_VN(0.161) |
| ɛ | ɛ ː _FR(0.333) | e_FR(0.421) |
| ʃ | ʃ_FR(0.56) | s_VN,ʃ_FR (0.121) |
| ə | ə_FR(0.098) | ø_FR(0.092) |
| ɛ̃ | ɛ̃_FR(0.25) | ɛχ_VN(0.284) |
| ɛ ː | œ_FR(0.139) | ɛ ː _FR(0.257) |
| f | f_FR(0.477) | f_VN(0.306) |
| g | v_EN,g_FR (0.143) | ŋ_VN(0.095) |
| h | ??? | ??? |
| ɥ | ɥ_FR(0.333) | ɥ_FR(0.3) |
| i | y_FR(0.245) | i_FR(0.263) |
| j | j_FR(0.523) | j_FR(0.296) |
| k | v_FR(0.165) | k_VN(0.119) |
| l | l_FR(0.135) | l_VN(0.137) |
| m | m_EN(0.441) | m_VN(0.361) |
| n | m_FR(0.137) | n_VN(0.171) |
| ɲ | ɲ_FR(0.462) | ɲ_FR(0.409) |
| ŋ | ??? | ??? |
| o | o_FR(0.2) | o_FR(0.237) |
| ø | ə_EN(0.333) | ɔ_FR(0.19) |
| oɔ | o_FR(0.167) | u_FR(0.333) |
| œ | o_EN(0.583) | ɤ_VN(0.315) |
| œ̃ | ɑ_FR(0.087) | a_VNŋ_VN (0.059) |
| øœ | ??? | ??? |
| p | v_EN(0.14) | v_VN(0.162) |
| ʁ | ʁ_FR(0.151) | ʁ_FR(0.055) |
| s | s_FR(0.352) | s_FR(0.403) |
| t | t_FR(0.086) | tʰ_VN(0.086) |
| u | u_FR(0.436) | u_FR(0.311) |
| v | v_FR(0.698) | v_FR(0.432) |
| w | w_FR,ɔ_EN (0.273) | w_FR(0.243) |
| y | y_FR(0.646) | y_FR(0.18) |
| z | z_FR(0.659) | z_FR(0.528) |
| ʒ | ʒ_FR(0.633) | ʒ_FR(0.198) |
| ∑= 42 | ∑phoneFR=32<br>∑phoneEN=7<br>∑NonAlign=3 | ∑phoneFR=21<br>∑phoneVN=18<br>∑NonAlign=2 |

b.

| EN | ENfr | ENvn |
|---|---|---|
| a | o_EN(0.255) | a_VN(0.483) |
| a ː | ??? | ??? |
| æ | æ_EN(0.151) | a_VN(0.182) |
| aɪ | aɪ_EN(0.371) | a_VN(0.219) |
| b | b_FR(0.283) | n_VN(0.13) |
| ɔ | o_EN(0.211) | ɔ_VN(0.139) |
| ɔɪ | ɔɪ_EN(0.306) | a ː _EN(0.6) |
| d | d_FR(0.161) | ŋ_VN(0.08) |
| ð | z_FR(0.274) | ð_EN(0.248) |
| dʒ | ʒ_FR,dʒ_EN (0.389) | ʐ_VN,s_EN (0.2) |
| e | e_EN(0.333) | e_EN(0.28) |
| ɛ | e_EN(0.166) | ɛ_EN(0.29) |
| ʃ | ʃ_FR(0.415) | ʃ_EN(0.545) |
| ə | œ_FR(0.103) | ɤ_VN(0.224) |
| f | f_EN(0.656) | f_VN(0.568) |
| g | g_FR(0.333) | v_EN,ŋ_VN (0.333) |
| h | h_EN(0.458) | h_EN(0.326) |
| i | i_EN(0.34) | i_EN(0.218) |
| ɪ | ɪ_EN(0.267) | ɪ_EN(0.322) |
| j | j_EN(0.323) | j_EN(0.261) |
| k | g_FR(0.167) | tʰ_VN(0.098) |
| l | l_FR(0.151) | l_VN(0.097) |
| m | m_EN(0.234) | ŋ_VN(0.182) |
| n | n_EN(0.268) | n_VN(0.276) |
| ŋ | ŋ_EN(0.523) | ŋ_EN(0.311) |
| o | o_EN(0.323) | ɤ_VN(0.281) |
| p | v_EN(0.213) | v_VN(0.31) |
| r | r_EN(0.096) | r_EN(0.098) |
| s | s_EN(0.281) | s_EN(0.353) |
| t | t_EN(0.084) | tʰ_VN(0.06) |
| tʃ | ʃ_EN(0.207) | s_EN(0.188) |
| u | u_EN(0.371) | u_VN(0.169) |
| ʊ | ø_FR(0.149) | ʌ_EN(0.28) |
| v | v_FR(0.308) | p_VN(0.21) |
| ʌ | ʌ_EN(0.128) | ʌ_EN(0.14) |
| w | w_EN(0.261) | w_EN(0.421) |
| z | z_FR(0.505) | s_EN(0.133) |
| ʒ | ??? | z_EN(0.6) |
| θ | θ_EN(0.283) | tʰ_VN,θ_EN (0.182) |
| ∑= 39 | ∑phoneEN=26<br>∑phoneFR=12<br>∑NonAlign=1 | ∑phoneEN=21<br>∑phoneVN=20<br>∑NonAlign=1 |

c.

| VN | VNfr | VNen |
|---|---|---|
| a | a_VN(0.227) | æ_EN(0.152) |
| aχ | aχ_VN(0.168) | e_EN,a_EN (0.17) |
| b | b_VN(0.219) | v_VN(0.333) |
| c | ??? | ??? |
| ɔ | o_FR(0.12) | o_EN(0.207) |
| cʰ | c_VN(0.19) | tʃ_EN(0.25) |
| ɔχ | ɔ̃_FR(0.187) | o_EN(0.526) |
| d | b_FR(0.109) | v_VN(0.312) |
| e | e_FR(0.29) | e_EN(0.368) |
| ɛ | ɛ ː _FR(0.311) | ɪ_EN(0.192) |
| ɛχ | a_VN(0.329) | ɛχ_VN(0.286) |
| f | f_FR(0.561) | f_EN(0.63) |
| ɣ | g_FR(0.407) | ɣ_VN(0.333) |
| ɤ | ɤ_VN(0.246) | ɤ_VN(0.222) |
| ɤχ | a ː _FR(0.085) | ʌ_EN(0.153) |
| h | h_VN(0.219) | h_EN(0.442) |
| i | i_FR(0.398) | i_EN(0.42) |
| ie | i_FR(0.123) | i_EN(0.189) |
| j | j_VN(0.138) | i_EN(0.096) |
| k | χ_VN(0.098) | χ_VN(0.188) |
| l | l_FR(0.175) | l_VN(0.149) |
| m | m_VN(0.17) | m_EN(0.527) |
| ɯ | u_FR(0.209) | ɯ_VN(0.267) |
| ɯɤ | ɯɤ_VN(0.126) | o_EN(0.194) |
| n | n_FR(0.208) | m_EN(0.253) |
| ɲ | ɲ_FR(0.162) | m_EN(0.171) |
| ŋ | ŋ_VN(0.181) | ŋ_VN(0.221) |
| o | ɑ̃_FR(0.18) | o_EN(0.213) |
| p | p_FR(0.133) | ŋ_VN(0.205) |
| s | ??? | ??? |
| ʂ | s_FR(0.416) | s_EN(0.276) |
| t | tʰ_VN(0.116) | v_VN(0.215) |
| tʰ | tʰ_VN(0.168) | tʰ_VN(0.291) |
| u | u_FR(0.434) | u_VN(0.493) |
| uo | u_FR(0.273) | uo_VN(0.147) |
| v | v_FR(0.32) | v_EN(0.363) |
| w | w_VN(0.09) | o_EN(0.112) |
| z | z_FR(0.427) | z_EN(0.613) |
| ʐ | z_FR(0.269) | z_EN(0.75) |
| χ | χ_VN(0.158) | χ_VN(0.241) |
| ∑= 40 | ∑phoneVN=17<br>∑phoneFR=22<br>∑NonAlign=2 | ∑phoneVN=15<br>∑phoneFR=24<br>∑NonAlign=2 |

*Tableau 2.5 : Tableau de prononciation des locuteurs non natifs (selon la matrice de confusion) et le symbole « ??? » signifie qu'aucun phonème correspondant au phonème de la première colonne du tableau n'est trouvé*

Avec les résultats des confusions de phonèmes des locuteurs non natifs présentés dans le tableau 2.5, nous pouvons conclure que :

- les locuteurs français empruntent beaucoup de phonèmes de leur langue maternelle dans le discours vietnamien ; par contre, ils arrivent à prononcer beaucoup plus de phonèmes anglais que les locuteurs vietnamiens dans leurs discours anglais. La même observation de substitution des phonèmes source/cible pour les locuteurs anglais peut être faite quand ils parlent le vietnamien et le français. D'un autre côté, les locuteurs vietnamiens utilisent beaucoup de phonèmes vietnamiens dans leur discours anglais et français. Cela met en évidence que le nombre de substitutions phonétiques dépend de la distance entre la langue cible et la langue source : le nombre de substitutions phonétiques est faible si les deux langues (source et cible) sont proches. Cette conclusion va dans le sens de l'analyse à partir du tableau API (Section 2.5.1). Par ailleurs, cela ouvre une voie pour tenter d'identifier la relation de distance entre plusieurs langues ou pour des langues inconnues (par exemple les langues peu dotées) en utilisant la matrice de confusion de phonèmes.
- Dans le cas où il y a substitution de phonèmes, les locuteurs non natifs substituent les phonèmes de langue cible par les phonèmes de la langue source qui sont les plus proches (selon l'API). Pour les phonèmes de la langue cible qui n'existent pas dans la langue source, les phonèmes de la langue source ou cible qui sont dans le même lieu d'articulation ou mode d'articulation (selon le tableau d'articulation de l'API) que les phonèmes de la langue cible sont souvent utilisés par les locuteurs non natifs. Par exemple, un locuteur vietnamien parlant français peut remplacer /ʃ/ et /g/ par les phonèmes /s/ et /ŋ/ respectivement (voir la figure 2.3).

Avec les résultats de la matrice de confusion de phonèmes présentés dans le tableau 2.5, nous pouvons également extraire les informations relatives aux phonèmes spécifiques des trois langues en question. L'objectif est de savoir quelles sont les phonèmes rares d'une langue cible que les locuteurs non natifs n'arrivent pas à prononcer. Par exemple, selon le tableau 2.6, les locuteurs français n'arrivent pas à prononcer le phonème vietnamien /ɤ/ alors que les locuteurs anglais peuvent le faire. Mais tous les locuteurs français et anglais ne peuvent pas prononcer les phonèmes vietnamiens /d, ɛ, ɣ, k, ɯ, ɔ, ʂ, t, ʐ/.

| | **Locuteur EN** | **Locuteur FR** | **Locuteur VN** | **Phonème Particulier d'une Langue** |
|---|---|---|---|---|
| **Langue EN** | NON | aː_EN, ɔ_EN, a_EN, ð_EN, ɛ_EN, p_EN | æ_EN, b_EN, d_EN, g_EN, dʒ_EN, m_EN, o_EN, t_EN, ʒ_EN | ə_EN,ɪ_EN,k_EN,ʊ_EN |
| **Langue FR** | e_FR, i_FR, k_FR, n_FR, ŋ_FR | NON | œ_FR, b_FR, ɛ_FR, g_FR, ɥ_FR, ɛ̃_FR, ʁ_FR, t_FR | ø_FR, h_FR, p_FR |
| **Langue VN** | b_VN, c_VN, j_VN, ɲ_VN, n_VN, p_VN, w_VN | ɤ_VN | NON | d_VN, ɛ_VN, ɣ_VN, k_VN, ɯ_VN, ɔ_VN, ʂ_VN, t_VN, ʐ_VN |

*Tableau 2.6: Les phonèmes des langues française, anglaise et vietnamienne qui ne sont pas retrouvés par les locuteurs non natifs (selon la méthode de la matrice de confusion)*

## 2.6. Conclusion

Dans ce chapitre, nous avons présenté notre corpus « MICA-MultiMeet » contenant de la parole native et non native pour quatre langues (français, anglais, khmer et vietnamien). Le corpus contient environ six heures de signal de parole (cinq heures de signal transcrit et une heure de signal non transcrit). La parole non native représente une grande partie du corpus (66 % du signal transcrit sont de la parole non native).

Un corpus de test a été extrait à partir du corpus précédent pour nos prochaines expérimentations. À cause de l'indisponibilité de la parole non native du khmer, le corpus de test ne contient que de la parole native et non native en français, anglais et vietnamien.

En outre, nous avons étudié les confusions de phonèmes faites par les locuteurs non natifs en obtenant de façon automatique une matrice de confusion de phonèmes de langues source et cible. Les résultats montrent qu'une telle méthode automatique permet de retrouver des phénomènes bien connus par les phonéticiens sur la parole non native (remplacement de phonèmes cibles par le phonème source le plus proche par exemple).

# Chapitre 3
# Observateur de langues

## 3.1. Introduction

Dans le chapitre précédent, nous avons présenté le corpus multilingue que nous allons utiliser lors du processus d'adaptation autonome de modèle acoustique multilingue. Nous rappelons que ce processus d'adaptation, qui réadapte le modèle acoustique lui-même lors du décodage, de manière non supervisée (sans connaître a priori la langue parlée ni l'origine du locuteur) et sans besoin de données d'adaptation, fonctionne selon un décodage acoustico-phonétique en deux passes (voir figure 1.1 du chapitre 1). La première passe détermine les caractéristiques linguistiques de la phrase décodée (l'observateur de langue) et la deuxième passe consiste à adapter le modèle acoustique multilingue en utilisant des connaissances extraites par le module nommé *observateur de langues* (OL).

Dans ce chapitre, nous présentons en détail le module *observateur de langues*. La définition de l'observateur de langues (OL) et les caractéristiques fournies par l'OL sont données dans la section 3.2. Deux approches différentes pour construire cet observateur sont présentées dans les sections 3.3 et 3.4, et les résultats d'expérimentation de ces deux approches sont détaillés dans la section 3.5.

## 3.2. Définition

Dans le processus d'adaptation non supervisé des modèles acoustiques multilingues, l'observateur de langues (OL) est un module cœur qui détermine les caractéristiques, liées à la langue parlée, d'un ou plusieurs segments inconnus de parole à décoder, afin que le module d'adaptation puisse décider comment adapter le modèle acoustique multilingue pour mieux décoder ces segments inconnus. Par « inconnus », nous entendons des segments pour lesquels nous ne savons pas à l'avance quelle langue est parlée et quelle est l'origine du locuteur (natif ou non natif).

D'après [Flege, 1997; Tan, 2007], les locuteurs non natifs empruntent des caractéristiques acoustiques de la langue maternelle (langue source ou L1) dans leurs discours de parole non

native (langue cible ou L2). Dans notre contexte d'adaptation non supervisée des modèles acoustiques multilingues, les caractéristiques utiles (langue d'origine L1 et langue parlée L2) d'un segment de la parole peuvent être obtenues par les scores postérieurs des langues $P(L_i)$. Par exemple, pour un segment fourni en entrée, si l'observateur de langue donne les scores postérieurs P (EN) = 0,5 ; P (FR) = 0,4 et P (VN) = 0,1 ; ce segment de parole pourra être considéré comme de l'anglais prononcé par un Français (ou inversement). Et si P (EN) = 0,8 ; P (FR) = 0,1 et P (VN) = 0,1 ; alors le segment de la parole sera probablement de l'anglais prononcé par un locuteur natif anglais.

Pour générer les scores postérieurs des langues, deux approches sont proposées : la première approche, appelée *language label voting* (LLV), se base sur les modélisations acoustiques multilingues, tandis que la deuxième est fondée sur les modélisations acoustiques, ainsi que sur des modèles phonotactiques, largement utilisés dans le domaine de la reconnaissance des langues (LID). Cette seconde approche sera nommée *Phone Recognizer following by Vector Space Modeling* (PR-VSM). Nous détaillons ces approches dans les deux sections suivantes.

## 3.3. *Language Label Voting (LLV)* : un observateur de langues à base de modélisations acoustiques multilingues

### 3.3.1. Définition

Pour générer les scores postérieurs des N langues en question, le processus LLV se décompose en deux modules principaux - le décodeur acoustico-phonétique multilingue et le calcul des scores postérieurs de langue - comme illustré dans la figure 3.1.

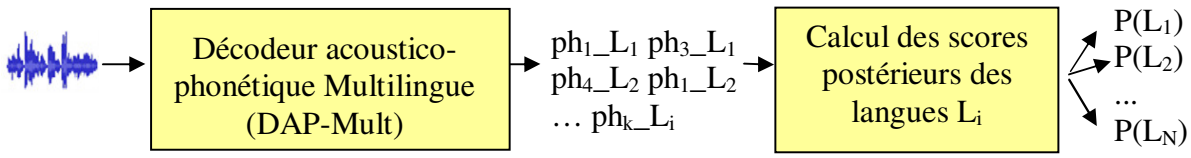



*Figure 3.1 : Observateur LLV*

Selon la figure 3.1, pour calculer les scores postérieurs de N langues (les langues parlées et les langues maternelles des locuteurs à traiter), nous proposons un décodeur acoustico-phonétique multilingue (DAP-Mult) avec les critères suivants :

- le modèle acoustique multilingue (MA-Mult) doit contenir toutes les unités acoustiques de ces N langues ; par exemple, ce modèle dispose d'unités

acoustiques de langue anglaise (EN), française (FR), mandarin (CN) et vietnamienne (VN), l'observateur LLV pourra, dans ce cas, être utilisé pour caractériser les segments de parole anglaise prononcé par des Français (ENfr) ou par des Chinois (ENcn) ou par des Vietnamiens (ENvn), et inversement ;

- chaque unité acoustique de MA-Mult doit être étiquetée par la langue à laquelle elle appartient ; par exemple, le phonème /a_EN/ est le phonème /a/ qui a été entraîné sur un corpus de langue anglaise prononcée par des locuteurs natifs anglais.

### 3.3.2. Décodeur acoustico-phonétique multilingue

D'après [Haton, 1991], un décodage acoustico-phonétique (DAP) est défini généralement comme une transformation de l'onde vocale en unités phonétiques. C'est donc une forme de transcodage qui fait passer d'un code acoustique à un code phonétique pour lequel le niveau de représentation passe du continu au discret.

Notre décodeur acoustico-phonétique multilingue (DAP-Mult) utilise uniquement un modèle acoustique multilingue (MA-Mult). Le modèle de langue (ML-Mult) du DAP-Mult est, dans ce cas, simplement un modèle plat où toutes les unités phonétiques des langues impliquées dans le modèle acoustique multilingue sont équiprobables. Cette approche est aussi appelée « *Flat Language Modelling*» ou « *Phone Loop Grammar Language Modeling* ».

#### *3.3.2.1. Choix de la méthode de combinaison des modèles acoustiques*

Pour créer un modèle acoustique multilingue, [Schultz, 2001] propose trois méthodes de combinaison de modèles acoustiques (ML-sep, ML-tag et ML-mix) détaillées dans la section 1.3.1 du chapitre 1. Au lieu de créer un modèle acoustique multilingue selon la méthode ML-tag et ML-mix, nous choisissons la méthode « séparation des langues » (ML-sep) dans notre tâche d'observation des langues, car la méthode de combinaison des modèles acoustiques ML-sep présente deux propriétés intéressantes :

- le processus de combinaison est plus rapide car les modèles acoustiques des langues traitées sont déjà disponibles ; la méthode ML-sep consiste à simplement fusionner les unités acoustiques des langues traitées sans réentraîner les modèles (aucune unité acoustique n'est partagée entre les langues) ;
- il est possible d'avoir, dans les hypothèses issues du DAP-Mult, des séquences de phonèmes étiquetés selon les langues auxquelles ils appartiennent, car chaque unité

acoustique dans une langue est apprise uniquement sur les données correspondantes de cette langue.

### *3.3.2.2. Choix des modèles acoustiques à combiner*

Rappelons que, dans notre corpus de test (section 2.4 du chapitre 2), en raison de l'indisponibilité de parole non native de la langue khmère, nous ne traitons que la parole native et non native de trois langues (EN, FR et VN), même si le corpus MICA-MultiMeet contient de la parole en quatre langues. Ainsi, nous ne combinons que les modèles acoustiques de l'anglais, du français et du vietnamien pour créer nos modèles acoustiques multilingues. Pour simplifier la combinaison des modèles acoustiques et aussi réduire la grande quantité d'états des HMMs dans le modèle acoustique multilingue, ce dernier est créé en combinant les modèles acoustiques monolingues indépendants du contexte (CI) et non pas dépendants du contexte (CD). Le tableau 3.1 présente les quantités de données d'entraînement des modèles acoustiques monolingues choisis.

| **Nom du corpus** | **#données (heure)** | **#locuteurs (natif)** | **#phonèmes** | **Critères d'entraînement choisis** |
|---|---|---|---|---|
| BREF120 [Lamel, 1991] | > 100 | 120 | 43 | - Contexte indépendant (CI)<br>- HMM à trois états (16 gaussiennes/état) |
| WSJ0 [Paul, 1992] | 20 | 140 | 40 | |
| VNSpeechCorpus [Le, 2004] | 20 | 50 | 41 | |

*Tableau 3.1 : Données d'entraînement des modèles acoustiques des langues traitées*

La figure 3.2 présente la combinaison ML-sep des modèles acoustiques de l'anglais (MA-EN), du français (MA-FR) et du vietnamien (MA-VN), pour le premier état du modèle de phonème /m/, noté « M-b ».

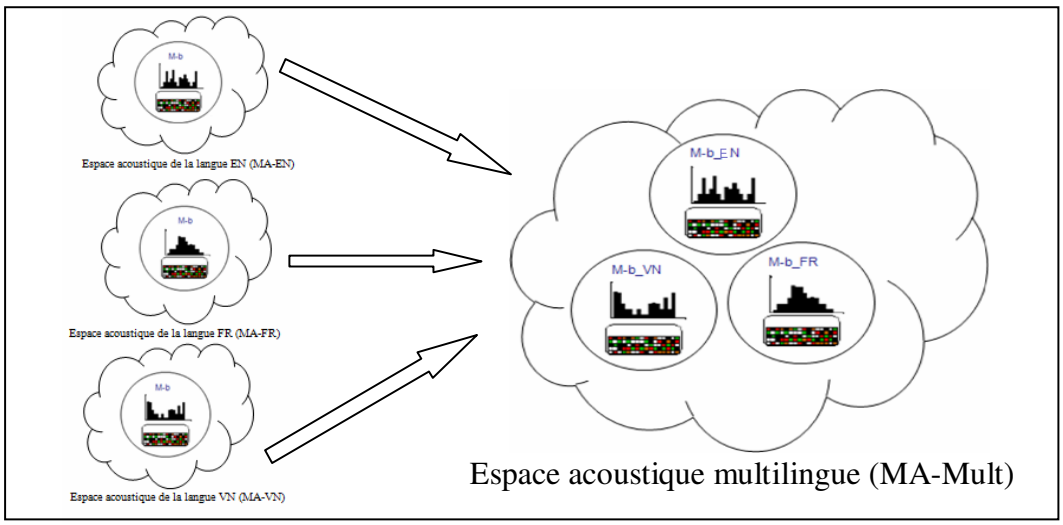



*Figure 3.2 : Espace acoustique - combinaison de trois modèles acoustiques (EN, FR, VN) selon la méthode ML-sep [Schultz, 2001]*

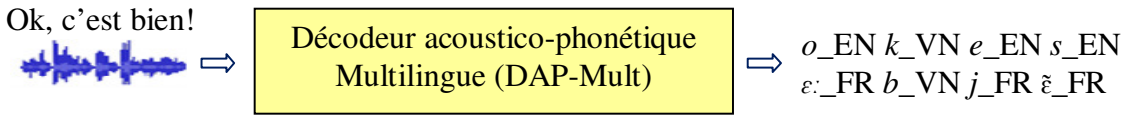



*Figure 3.3 : Exemple de décodage acoustico-phonétique multilingue en sortie de notre système*

La figure 3.3 illustre un exemple de notre décodeur acoustico-phonétique multilingue qui génère une séquence de phonèmes correspondant au segment de parole « Ok, c'est bien ! ».

Dans la figure 3.3, chaque phonème en sortie est étiqueté par la langue à laquelle il appartient. Notre concept d'adaptation autonome utilisant un observateur de langue est issu, en fait, de ce constat initial.

### 3.3.3. Génération des scores postérieurs

À partir de la séquence phonétique produite par le décodage acoustico-phonétique multilingue, nous proposons trois méthodes pour produire les probabilités a posteriori des langues en question. Ces trois méthodes sont détaillées dans les sections suivantes.

#### *3.3.3.1. Première méthode (Ph-EquiPro)*

La méthode par « phones équiprobables » (Ph-EquiPro) traite les phonèmes étiquetés dans la séquence phonétique fournie par DAP-Mult de manière équitable. Le lecteur trouvera une explication plus claire avec un exemple concret sur le calcul du score postérieur à base de la méthode Ph-EquiPro à la fin de cette section. La méthode Ph-EquiPro fournit les probabilités de langues en utilisant l'hypothèse du système acoustico-phonétique multilingue (figure 3.3) par la formule suivante :

$$P(L_i) = \frac{n(L_i)}{N} \quad , \sum_{i=1}^{3} P(L_i) = 1 \tag{3.1}$$

où $P$, $L_i$, $n$, $N$ représentent respectivement, le score postérieur, la langue (FR ou EN ou VN), le nombre de phonèmes d'une langue dans la séquence phonétique décodée, le nombre total de phonèmes trouvés dans cette séquence phonétique.

#### *3.3.3.2. Deuxième méthode (Ph-ProPart)*

Dans la méthode Ph-ProPart, nous ne considérons que les *phonèmes spécifiques* trouvés dans la séquence phonétique. Les phonèmes spécifiques sont ceux qui existent dans une seule langue. Par exemple, le phonème /ɑ̃/ est un des phonèmes spécifiques de la langue française (c'est une voyelle nasale), car il n'appartient ni à la langue anglaise ni à la langue vietnamienne qui ne possèdent pas de voyelles nasales. Le calcul du score postérieur d'une langue est obtenu par la formule suivante :

$$P(L_i) = \frac{n'(L_i)}{N'} \quad , \sum_{i=1}^{3} P(L_i) = 1 \tag{3.2}$$

où *n'* et *N'* signifient respectivement le nombre de phonèmes spécifiques d'une langue et le nombre total des phonèmes spécifiques trouvés dans la séquence phonétique.

#### *3.3.3.3. Troisième méthode (Ph-ProVar)*

Le score postérieur d'une langue est calculé, dans ce cas, à partir de la somme du nombre de phonèmes communs avec les autres langues et du nombre des phonèmes spécifiques trouvés dans la séquence phonétique. Le phonème commun est un phonème qui existe au moins dans deux langues. Par exemple, le phonème /a/ est un phonème commun aux trois langues considérées (FR, EN et VN).

Il est important de préciser qu'un phonème spécifique est modélisé par un seul HMM dans le modèle acoustique multilingue, alors qu'un phonème *commun* est modélisé par plusieurs HMMs différents. Par exemple, dans notre modèle acoustique multilingue qui contient tous les phonèmes de trois langues, il y a trois HMMs pour le phonème commun /a/ (un HMM pour chaque langue) mais il n'y a qu'un seul HMM pour le phonème /ɑ̃/ du français. Pour cette raison, dans la méthode de calcul des scores postérieurs Ph-ProVar, nous donnons un score plus important aux phonèmes spécifiques en multipliant leurs scores par le nombre de langues. Le calcul du score postérieur d'une langue est obtenu par la formule suivante :

$$P(L_i) = \frac{(n'(L_i) * N_L) + n_{commun}}{\sum_{i=1}^{N_L}((n'(L_i) * N_L) + n_{commun})} \quad , \sum_{i=1}^{3} P(L_i) = 1 \tag{3.3}$$

où $N_L$ et $n_{commun}$ désignent respectivement le nombre des langues traitées et le nombre de phonèmes communs qui existent dans la langue $L_i$ et dans les deux autres langues.

Par exemple, pour la séquence suivante, si « o_EN k_VN e_EN s_EN ɛː_FR b_VN j_FR ɛ̃_FR » est l'hypothèse du système acoustico-phonétique lors du décodage d'un segment « Ok, c'est bien ! » (figure 3.3), il y aurait six phonèmes communs (/o/, /k/, /e/, /s/, /b/ et /j/) pour toutes les trois langues traitées et deux phonèmes spécifiques du français (/ɛː/ et /ɛ̃/) (les informations concernant la liste des phonèmes communs et spécifiques par rapport aux langues traitées ont été présentées dans le tableau 2.5 du chapitre 2). Dans ce cas, le calcul des scores postérieurs à base des trois méthodes précédentes donne :

- Ph-EquiPro: *P(EN) = P(FR) = 3/8 = 0,375 ; P(VN) = 2/8 = 0,25* ;
- Ph-ProPart : *P(EN)=P(VN)=0 ; P(FR)=2/2=1 ;*

- Ph-ProVar : $n_{commun}$ *de EN, FR, VN = 6 ((/o/, /k/, /e/, /s/, /b/,/j/) ;* $N_L$*=3 ; n'(FR)=2 (/ɛ:/,/ɛ̃/) ; alors selon l'équation 3.3:*

  ➔ *nombre total des phonèmes =* $\sum_{i=1}^{N_L}((n'(L_i)*N_L)+n_{commun}) = 24$ ;

  ➔ *les scores postérieurs de l'anglais et du vietnamien sont : P(EN)=P(VN)=[1(o)+1(k)+1(e)+1(s)+1(b)+1(j)]/24=6/24=0,25 ;*

  ➔ *le score postérieur du français est : P(FR)= [ 3 (nombre de langues) * 2 (phonèmes spécifiques du français) + 6 (phonèmes communs avec deux autres langues) ] / nombre total de phonèmes = [3* (1(ɛ:)+1(ɛ̃)) + (1(o)+1(k)+1(e)+1(s)+ 1(b)+1(j))]/24 = [3*(2)+(6)]/24=12/24 = 0,5 ;*

### 3.3.4. Première évaluation des méthodes proposées

Dans un premier temps, nous avons voulu vérifier si les scores postérieurs générés sont utilisables pour un processus d'adaptation autonome.

Pour les valider, nous avons extrait 100 segments consécutifs de parole à partir de notre corpus de réunions « MICA-MultiMeet » (avec 60 secondes par segment). Dans ce cas, un segment de flux audio peut avoir plusieurs locuteurs et plusieurs langues parlées (y compris la langue khmère). Ces segments sont différents des segments de test que nous avons présentés dans la section 2.4 (un segment contient seulement une des trois langues parlées : EN, FR et VN). Pour traiter ces segments de flux audio, nous utilisons un modèle acoustique quadrilingue pour le décodage acoustico-phonétique au lieu d'un modèle acoustique trilingue (EN, FR et VN). Ce modèle, dans ce cas, est simplement créé en combinant les trois modèles acoustiques des trois langues, mentionnés dans la section 3.3.2.1, avec un modèle acoustique du khmer (MA-KH) développé par [Seng, 2008a].

La figure 3.4 illustre un exemple de calcul des scores postérieurs pour un segment de 60 seconds de flux audio extrait, selon les trois approches de génération de scores postérieurs.

La première ligne de la figure 3.4 présente la référence d'un segment de flux audio de 60 secondes. Cette référence indique quelle langue est parlée sur quelle portion, ainsi que l'origine du locuteur. Par exemple, le symbole « *EN-fr 35.5s-38.5s* » signifie que ce segment est en langue anglaise (EN), énoncé par un locuteur français (fr), et débute à l'instant t = 35,5 secondes et finit à l'instant t = 38,5 secondes. La 2ème ligne présente les blocs d'analyse automatique obtenus toutes les 10 secondes avec notre observateur de langues. Les lignes 3, 4 et 5 présentent les scores postérieurs (correspondant à chaque bloc de la 2ème ligne) calculés par les méthodes Ph-EquiPro, Ph-ProPart, Ph-ProVar respectivement.

Les mots clés OK, NON et N/A indiquent la justesse de ces résultats pour les trois méthodes en les comparant avec la référence (ligne 1). N/A signifie que, dans le segment

traité, il n'y a pas de phonème spécifique, pour aucune des quatre langues. N/A n'existe que dans le résultat de Ph-ProPart quand il n'y a pas de phonème spécifique dans le bloc traité.

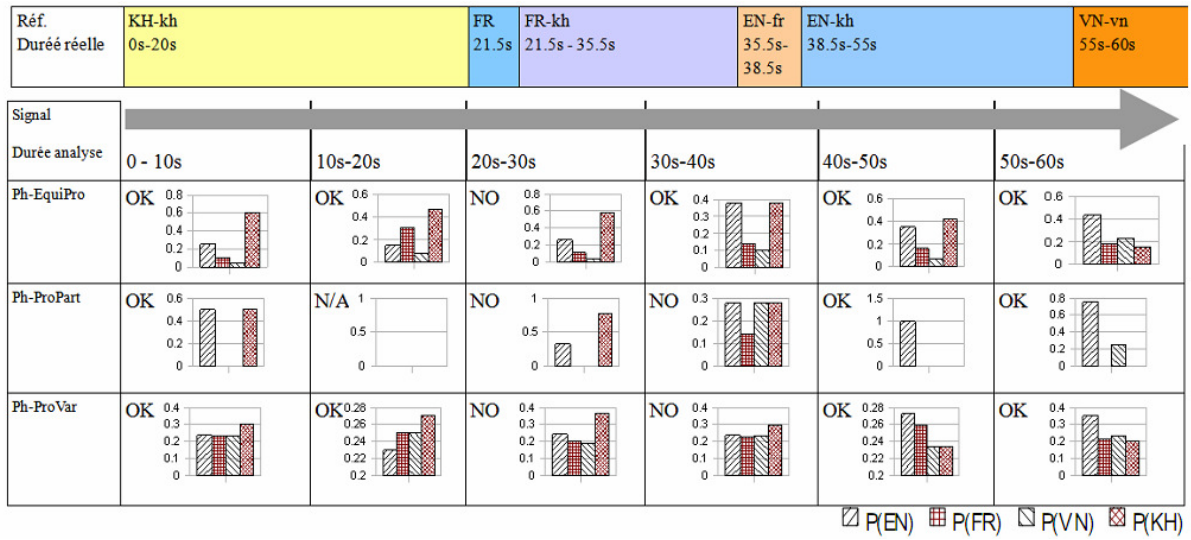



*Figure 3.4 : Exemple de comportement de l'observateur de langues selon l'approche LLV*

Selon la figure 3.4, nous pouvons faire les commentaires suivants sur les trois méthodes.

- La méthode Ph-ProPart fournit les scores postérieurs les moins intéressants, car elle ne donne aucun score postérieur s'il n'y a pas de phonème spécifique dans le segment traité (le cas N/A);
- Même si les méthodes Ph-ProVar et Ph-EquiPro donnent des résultats d'observation corrects, les scores postérieurs de la méthode Ph-EquiPro semblent plus clairs que ceux de la méthode Ph-ProVar. Nous pouvons remarquer, par exemple, que pour les deux premiers blocs (0 s à 20 s) qui représentent les segments de la parole native KH, le score postérieur du KH fourni par Ph-EquiPro est nettement plus élevé que les scores des autres langues, tandis que Ph-ProVar donne aussi des scores importants aux autres langues, car Ph-ProVar augmente les scores des autres langues s'il trouve des phonèmes KH qui sont partagés avec les autres langues (des phonèmes communs).
- Souvent, les méthodes Ph-EquiPro et Ph-ProVar ne donnent pas de bons scores postérieurs des langues quand un segment d'analyse contient de la parole en plusieurs langues, parlées par plusieurs locuteurs. Par exemple, dans le bloc de 20 s à 30 s, Ph-EquiPro fournit des scores postérieurs importants aux langues khmère et anglaise bien que le segment de référence soit constitué de parole native et non native française parlée par un Cambodgien ; dans le bloc de 30 s à 40 s, le score postérieur du français est plus petit que ceux de l'anglais et du khmer alors

que ce bloc contient une grande partie en langue française ; même observation pour le bloc de 50 s à 60 s.

Par ailleurs, si nous observons les résultats produits par la méthode Ph-EquiPro dans des blocs contenant une seule langue parlée, nous trouvons que :

- dans les deux premiers blocs (0 s à 20 s), le score postérieur du khmer est plus important que les autres. Cela est correct, car ces blocs correspondent à de la parole en langue khmère ;
- on peut faire la même observation pour le bloc de 40 s à 50 s qui est de la parole non native anglaise prononcée par un Cambodgien, le score postérieur du khmer et celui de l'anglais sont plus grands que ceux des deux autres langues ;

Après avoir observé les résultats de 100 segments [Sam, 2009], on constate que les méthodes Ph-EquiPro et Ph-ProVar donnent les résultats d'observation très compétitifs et meilleurs que ceux de la méthode Ph-ProPar.

Selon les résultats d'analyse précédents, il est envisageable que les scores postérieurs générés par LLV (surtout par la méthode Ph-EquiPro) soient utilisés dans le processus d'adaptation non supervisée du modèle acoustique multilingue surtout avec les segments dans notre corpus de test, car chaque segment de parole du corpus contient seulement une seule langue (une des trois langues traitées) parlée par un seul locuteur (natif ou non natif).

Il est nécessaire de préciser que le modèle acoustique quadrilingue (EN, FR, VN et KH) n'est utilisé que dans cette section, car un segment de flux audio peut contenir quatre langues parlées. Dans les autres tâches d'adaptation, nous n'utilisons que le modèle acoustique trilingue (EN, FR et VN).

## 3.4. PR-VSM : un observateur de langue fondé sur une approche phonotactique multilingue

Comme alternative à l'approche LLV qui est une approche simple fondée sur la sortie du système acoustico-phonétique, nous étudions également un observateur de langues qui se base non seulement sur le décodage acoustico-phonétique, mais aussi sur un modèle phonotactique. Ce type d'observateur de langue s'appelle PR-VSM (*Phone Recognize following by Vector Space Modeling*) et a été étudié par [Li, 2007] dans son travail sur la reconnaissance des langues (LID). L'observateur PR-VSM se compose de deux parties : la première partie (front-end) est un module de reconnaissance acoustico-phonétique (PR) et la modélisation de

l'espace vectoriel (VSM) selon l'approche phonotactique concerne la deuxième partie. Dans la première partie du PR-VSM, nous pouvons utiliser soit plusieurs décodeurs acoustico-phonétiques (un décodeur par langue) soit un seul décodeur acoustico-phonétique appelé « décodeur universel » (*Universal Phone Recognizer*, UPR).

Dans notre travail, nous utilisons le système acoustico-phonétique multilingue déjà présenté dans la figure 3.3 comme décodeur universel pour l'approche PR-VSM. La figure 3.5 illustre le processus complet de l'approche PR-VSM .

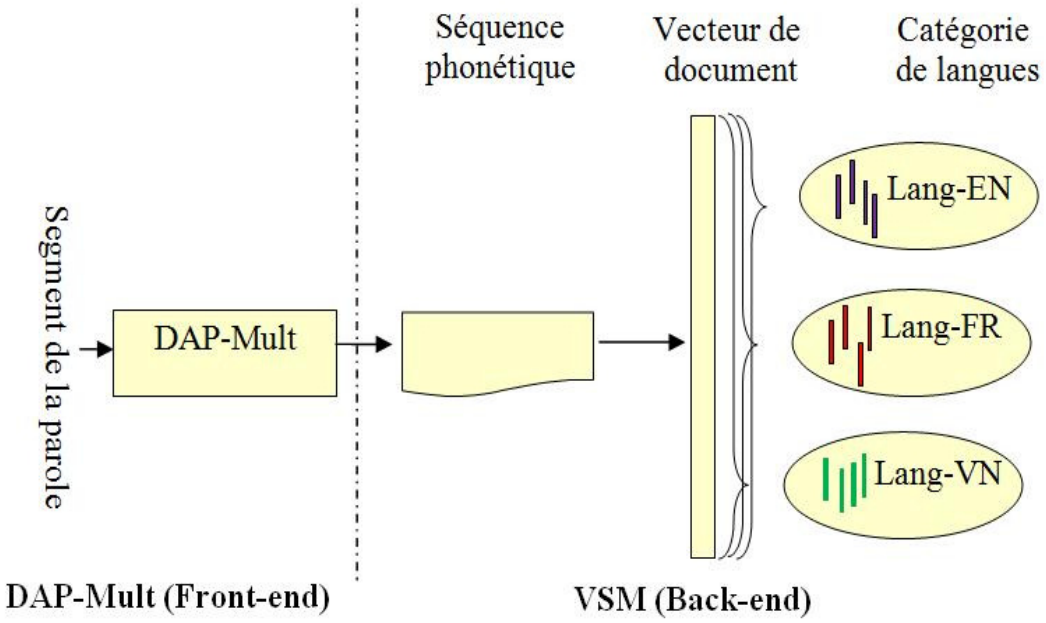



*Figure 3.5 : Schéma de l'observateur selon l'approche PR-VSM : DAP-Mult (front-end) suivi de VSM (back-end)*

Pour générer les scores postérieurs, l'observateur de langues PR-VSM a besoin de segments de parole des langues traitées pour entraîner son modèle VSM. Pour cela, nous avons extrait six heures de segments de parole native en anglais, français et vietnamien (deux heures pour chaque langue) à partir des corpus WSJ0 [Paul, 1992], BREF120 [Lamel, 1991] et VNSpeechCorpus [Le, 2004].

Ces segments de parole, avec leur étiquette de langue, sont décodés par notre décodeur acoustico-phonétique multilingue (DAP-Mult, figure 3.3) et les sorties du décodeur (les séquences phonétiques) sont converties en une collection de vecteurs « sonores » (vecteurs documentaires) avant d'être classés avec un SVM (*Support Vector Machines*) [Sebastiani, 2002]. En somme, cette approche est similaire aux « sacs de mots » utilisés en catégorisation de textes mais avec importance de l'ordre des symboles dans chaque vecteur (*bag-of-sound phonotactic approach*). Plus de détails sur l'approche « sacs de sons » sont disponibles dans [Ma, 2005].

Lors de la phase de test (nous utilisons les données de test mentionnées dans la section 2.4 du chapitre 2), un segment de parole inconnu est converti en une requête sous forme de vecteur documentaire (le processus de transformation suit exactement la même procédure que lors de l'apprentissage, de sorte que les scores postérieurs de la langue soient générés comme dans le cas de la classification des documents textuels [Joachims, 2002].

En résumé, un observateur de langues fondé sur l'approche PR-VSM génère les scores postérieurs des langues traitées par la formule suivante :

$$P(L_i) = \frac{LogP(T \mid VSM(L_i))}{\sum_{j=1}^{N_L} LogP(T \mid VSM(L_j))} \tag{3.4}$$

où T est la séquence de phonèmes fournis par notre décodeur acoustico-phonétique multilingue (DAP-Mult).

## 3.5. Résultats d'expérimentation

### 3.5.1. Évaluation de la performance de l'observateur de langues

Nous proposons deux métriques pour évaluer la performance de l'observateur de langues.

- Métrique LID : dans ce cas, la performance de l'observateur est définie exactement comme celle de l'identification des langues (LID). Cela signifie que seulement le meilleur score postérieur généré par l'observateur est utilisé pour choisir la langue la plus probable.
- Métrique LID+ORG : nous proposons également une autre mesure légèrement différente de la précédente. Cette fois, un test est considéré comme réussi si le maximum des scores postérieurs fournis par l'observateur correspond soit à la langue parlée soit à la langue maternelle du locuteur (dans ce dernier cas, le deuxième meilleur score postérieur doit être celui de la langue parlée, sinon on considère que le test a échoué). Par exemple, si les scores postérieurs sont P (FR) = 0,5 ; P (EN) = 0,4 ; P (VN) = 0,1 ; et si la référence correspond à la langue anglaise parlée par un locuteur français, alors la métrique LID indique une réponse incorrecte d'observation, tandis que la métrique LID+ORG indique une bonne observation. On remarque que l'erreur LID+ORG est une borne inférieure de l'erreur LID simple.

À partir des résultats d'évaluation des performances présentées dans le tableau 3.2, nous constatons que, parmi les trois observateurs de langues LLV à base du système acoustico-phonétique multilingue, la méthode nommée « phones équiprobables » (Ph-EquiPro) donne la meilleure performance d'observation des langues selon le taux d'erreur moyen (dernière ligne du tableau 3.2). Par ailleurs, la performance des méthodes LLV selon la métrique LID n'est pas très satisfaisante, car le taux d'erreur d'observation moyen est très élevé (39 % pour Ph-EquiPro, 56 % pour Ph-ProPart et 49 % pour Ph-ProVar).

| **Parole** | **LLV : Ph-EquiPro** | **LLV : Ph-ProPart** | **LLV : Ph-ProVar** | **PR-VSM** |
|---|---|---|---|---|
| ENen | 59/59 | 88/88 | 78/78 | **6/6** |
| ENfr | 21/15 | 96/89 | 66/51 | **17/0** |
| ENvn | 85/39 | 94/58 | 91/48 | **9/9** |
| FRfr | 18/18 | 34/34 | 24/24 | **0/0** |
| FRen | **0/0** | 5/5 | 5/0 | 2/**0** |
| FRvn | 53/**20** | 89/**20** | 73/24 | **24/**24 |
| VNvn | 5/5 | **0/0** | **0/0** | 26/26 |
| VNen | 77/70 | **23/0** | 50/0 | 50/29 |
| VNfr | 43/**2** | 38/20 | **36/**50 | 57/30 |
| Moyenne | 39/23 | 56/36 | 49/30 | **19/12** |

*Tableau 3.2 : Performances des observateurs de langue proposés (en %) : les chiffres à gauche de la barre oblique « / » représentent le taux d'erreur d'observation selon la métrique LID tandis que ceux à droite indiquent le taux d'erreur d'observation selon la métrique LID+ORG (les chiffres en gras représentent les meilleurs taux d'erreur d'observation)*

Cependant, selon la métrique LID+ORG qui prend en compte la langue d'origine du locuteur, le taux d'erreur d'observation diminue d'un facteur deux si nous comparons avec les résultats utilisant la métrique LID (23 % pour Ph-EquiPro, 36 % pour Ph-ProPart et 30 % pour Ph-ProVar). Cependant, il semble clair que les observateurs LLV, qui ne sont fondés que sur le système acoustico-phonétique, donnent une performance moins bonne que l'observateur fondé sur l'approche PR-VSM. Pour la suite de nos expériences, nous ne considérerons que l'observateur de langues PR-VSM qui semble le plus performant.

Selon les résultats d'observation des langues PR-VSM, nous pouvons conclure que :

- selon la métrique LID, les taux d'erreur pour la parole native sont généralement plus faibles que ceux pour la parole non native ; ce résultat, qui montre que l'identification des langues est plus difficile sur la parole non native, est cohérent avec une étude précédente de [Wanneroy, 1999].
- nous remarquons aussi que, dans l'évaluation de l'observateur selon la métrique LID+ORG, les différences de performance entre parole native et non native sont atténuées. Par exemple, pour la parole native et non native en langue vietnamienne, selon la métrique LID, les taux d'erreur de la parole VNvn, VNfr et VNen sont 26 %, 50 % et 57 % respectivement, mais selon l'évaluation LID+ORG les taux d'erreur de VNfr et VNen sont seulement 29 % et 30 % et sont très proches de ceux de la parole native VNvn (26 %). Cela signifie qu'il est possible d'utiliser les scores postérieurs fournis par l'approche PR-VSM pour capturer les informations de langue d'un segment de parole, non seulement dans le cas de la langue parlée, mais aussi dans celui de la langue d'origine du locuteur ; nous analysons ces scores postérieurs plus en détail dans la section suivante.

### 3.5.2. Analyse des scores postérieurs

Dans cette section, nous étudions des relations entre la parole native et non native des trois langues traitées EN, FR et VN, selon les scores postérieurs fournis par l'observateur de langue PR-VSM.

Nous présentons les scores postérieurs de tous les segments de test dans un espace à trois dimensions illustré par la figure 3.6. Chaque point dans cette figure représente un segment de parole extrait du corpus de test.

A partir de cette représentation de nos phrases dans cet espace à trois dimensions, nous pouvons dire que :

- les groupes des segments de parole native (ENen, FRfr et VNvn) restent nettement dans des zones différentes de l'espace ; mais les groupes ENen et FRfr semblent un peu plus proche du groupe VNvn (figure 3.7). Avec cette observation, il semble que PR-VSM soit capable de séparer les groupes de parole native (cela confirme les valeurs sur les lignes 1, 4 et 7 du tableau 3.2).

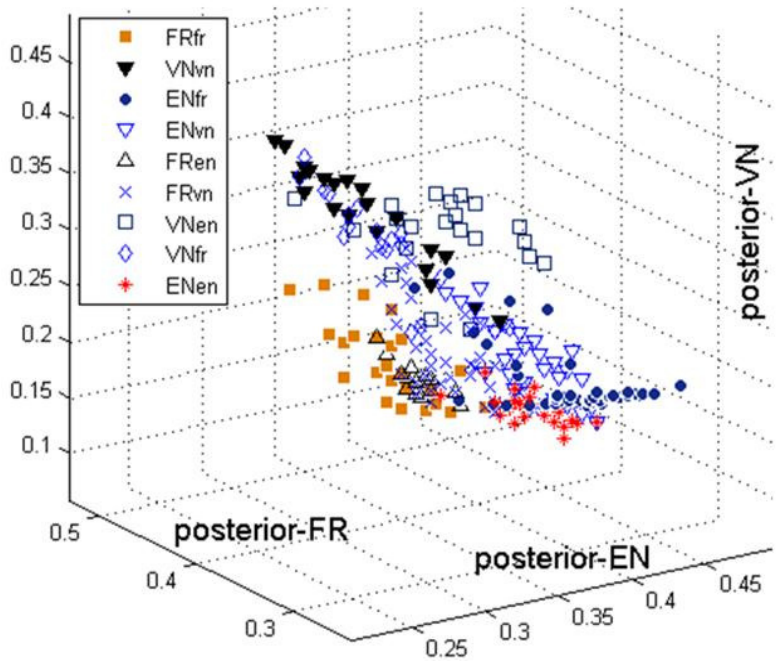



*Figure 3.6 : Localisation dans un espace à 3D des différents groupes de parole native et non native*

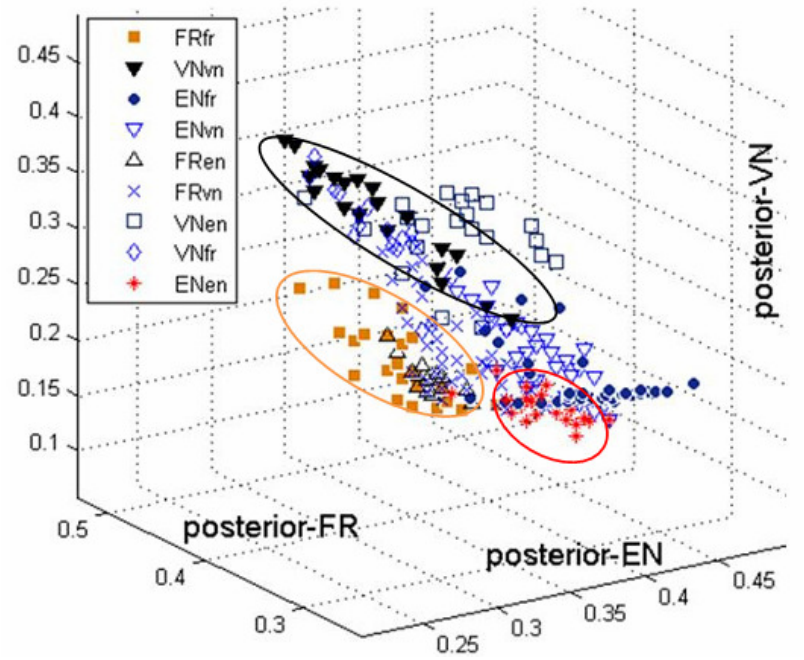
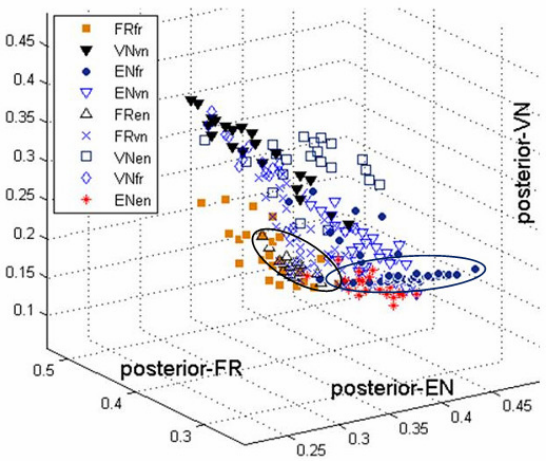



*Figure 3.7 : Observation des trois groupes de parole native sur l'espace à 3D*

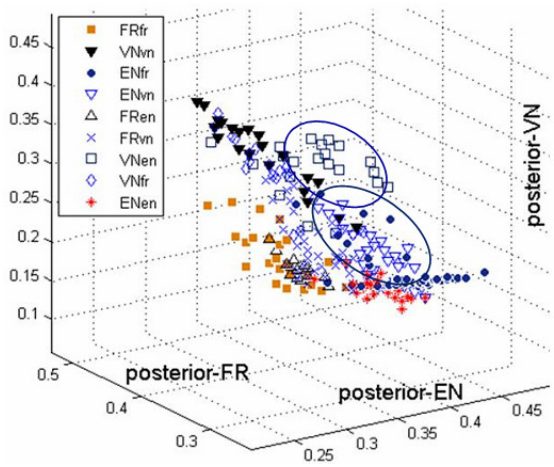
- la zone principale correspondant à un groupe de parole non native se situe généralement près de deux groupes de parole native : l'un correspond à la langue parlée et l'autre correspond à l'origine du locuteur (langue maternelle) ; par exemple, le groupe correspondant à la parole non native anglaise parlée par des locuteurs

français (ENfr) et celui de la parole non native française parlée par les locuteurs anglais (FRen) sont situés plus près des groupes de parole native ENen et FRfr que du groupe VNvn (figure 3.8a). Il s'avère alors que les scores postérieurs des langues générés par PR-VSM pourraient être aussi utilisés pour l'adaptation non supervisée du modèle acoustique multilingue pour différents groupes de parole non native (différentes origines des locuteurs).

a. ENfr vs. FRen




b. ENvn vs. VNen




c. FRvn vs. VNfr

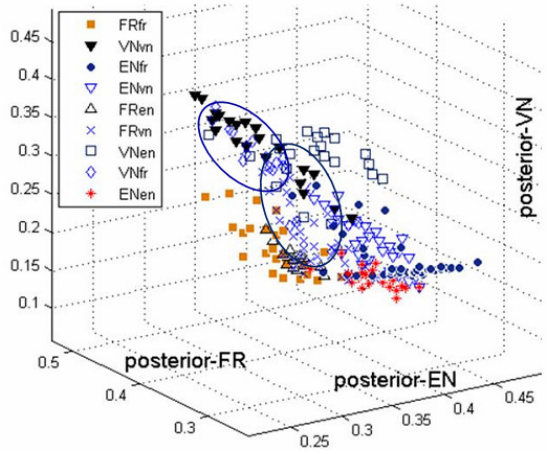



*Figure 3.8 : Observation des six groupes de parole non native sur l'espace à 3D*

## 3.6. Conclusion

Dans ce chapitre, nous avons présenté différentes approches pour construire un observateur de langues qui génère des scores postérieurs (pour chaque langue). Nous pensons qu'un tel observateur peut fournir des informations utiles pour l'adaptation non supervisée d'un modèle acoustique multilingue lors du décodage d'un segment de parole pour lequel la

langue et l'origine du locuteur sont inconnus. L'observateur de langue fondé sur l'approche PR-VSM semble le plus efficace et sera considéré désormais dans notre processus d'adaptation non supervisée décrit au chapitre suivant.

# Chapitre 4 :
# Adaptation des modèles acoustiques multilingues

## 4.1. Contexte d'adaptation

Nous rappelons que notre but consiste à résoudre deux défis principaux pour améliorer la performance du système de reconnaissance de la parole de type « réunion multilingue » : 1) les langues parlées et les langues d'origine des locuteurs (natifs ou non natifs) des segments de la parole à décoder sont inconnues ; 2) aucune donnée n'est disponible au départ pour l'adaptation des modèles acoustiques multilingues. Cela nous invite à étudier l'une des approches d'adaptation automatique non supervisée appelée *adaptation autonome* des modèles acoustiques multilingues.

L'adaptation autonome utilise un module appelé « observateur de langues » pour capturer des informations (les langues parlées et les langues d'origine des locuteurs) sur les segments de parole inconnus. Elle se base sur les scores postérieurs des langues générés à partir de l'observateur dans une première passe, puis, elle fait appel au module « adaptation des modèles acoustiques » pour adapter le modèle acoustique multilingue du décodeur de la première passe en fonction des connaissances fournies par l'observateur de langues et les résultats d'une première passe de décodage (aucune donnée d'adaptation supplémentaire n'est utilisée).

Dans ce chapitre, nous présentons en détail le module d'adaptation des modèles acoustiques multilingues. Le lecteur trouvera les détails des différentes techniques d'adaptation non supervisée « en ligne » et « hors ligne » (*online and offline unsupervised adaptation*) des modèles acoustiques dans la section 4.2. Les méthodes d'évaluation de la performance de ces techniques seront présentées dans la section 4.3. Les résultats d'expérimentation de toutes les techniques d'adaptation étudiées seront détaillés dans la section 4.4.

## 4.2. Différentes techniques d'adaptation non supervisée des modèles acoustiques multilingues

Nous étudions deux types d'adaptation non supervisée d'un modèle acoustique multilingue. Le premier type considère, dans son processus d'adaptation, le segment de la parole en cours de décodage et aussi les segments déjà décodés (les segments dans l'historique du décodage). Ce type d'adaptation est nommé « hors ligne » (*offline multilingual acoustic model adaptation*). Contrairement à l'adaptation « hors ligne », le deuxième type est l'adaptation « en ligne » (*online multilingual acoustic model adaptation*) et ne considère que le segment de parole en train d'être décodé.

Dans les sections suivantes, nous détaillons nos études sur trois techniques d'adaptation « en ligne » et deux techniques d'adaptation « hors ligne ».

### 4.2.1. Adaptation « en ligne » du modèle acoustique multilingue

#### *4.2.1.1. Maximum Likelihood Linear Regression (MLLR)*

Dans le chapitre 1, nous avons mentionné que les deux techniques d'adaptation des modèles acoustiques MLLR (*Maximum Likelihood Linear Regression*) et MAP (*Maximum a Posteriori*) sont le plus souvent utilisées pour créer des modèles adaptés au nouveau locuteur ou à l'environnement. Ces deux modèles, surtout le modèle MLLR, sont largement employés dans plusieurs domaines du traitement de la langue naturelle comme l'adaptation du modèle de locuteur pour la reconnaissance du locuteur et l'adaptation non supervisée des modèles acoustiques pour la reconnaissance automatique de la parole.

Dans la technique MLLR, les paramètres d'adaptation sont mis en commun et mis à jour avec des matrices de transformation, comme décrit dans la section suivante. L'approche MAP, quant à elle, propose une adaptation plus fine des modèles, mais nécessite beaucoup de données d'adaptation pour réestimer de façon fiable les paramètres pour tous les modèles utilisés dans le système. L'approche MLLR, au contraire, peut fournir une adaptation avec un nombre limité de segments d'adaptation.

Nous examinons dans cette section uniquement la technique d'adaptation MLLR, car nous considérons que notre application ne présente pas de données (les paroles natives et non natives de trois langues traitées) suffisantes pour l'adaptation MAP.

Il y a plusieurs types d'adaptation MLLR, comme *MLLR* sur les moyennes (*mean only*) [Leggetter, 1995], et sur moyennes et variances (*mean and variance*) [Gales, 1996]. [Ahn, 2000; McDonough, 1997] ont proposé, quant à eux, une variante nommée *MLLR to speaker-adaptive training* (MLLR-SAT). Selon des études précédentes sur ces trois types d'adaptation MLLR [Huang, 2001], MLLR-SAT donne la meilleure performance, mais son processus d'adaptation est très coûteux en temps. Par contre, le processus *MLLR on mean only* est le plus rapide et présente une performance très compétitive par rapport à celle de *MLLR on mean and variance*. Pour cette raison, toutes les adaptations MLLR mentionnées dans ce chapitre seront du type proposé dans [Leggetter, 1995] (*MLLR on mean only*).

L'approche de [Leggetter, 1995] consiste à trouver une matrice de transformation linéaire permettant de convertir le vecteur moyen d'une gaussienne, en un vecteur moyen adapté.

$$\mu' = \boldsymbol{A}^{(s)} \mu + \boldsymbol{b}^{(s)} \tag{4.1}$$

où μ' est un vecteur moyen adapté, μ est un vecteur moyen original (ou celui de l'itération précédente), **A** est une matrice de transformation de dimension n×n et **b** est un vecteur n-dimensionnel appelé le biais. Ici *n* est la dimension des vecteurs caractéristiques (*feature vector*), et ***(s)*** désigne un locuteur spécifique. La transformation **A** est calculée de manière à maximiser la vraisemblance des données d'adaptation par rapport au modèle selon un schéma EM (*Expectation/Maximization*) [Dempster, 1977].

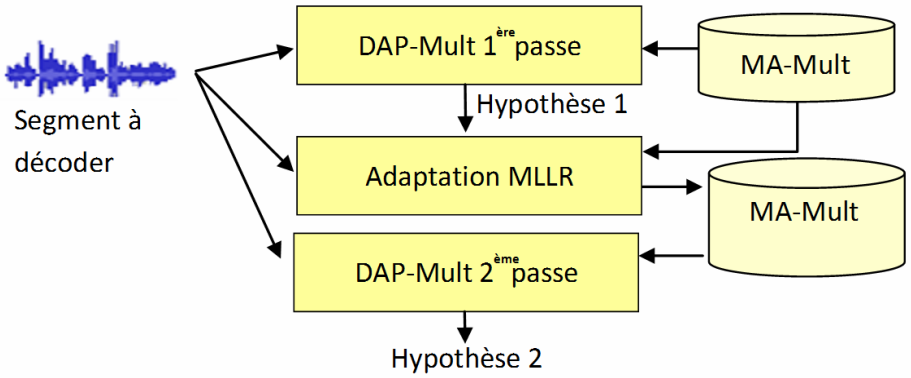



*Figure 4.1 : Processus d'adaptation « en ligne » non supervisé MLLR*

La figure 4.1 montre comment nous appliquons le processus d'adaptation MLLR en n'utilisant que le segment de la parole en cours du décodage. Dans nos études d'adaptation autonome, MLLR est le seul processus d'adaptation non supervisée qui ne dépend pas de l'observateur de langues (OL). Cette technique sera considérée comme une référence (*baseline*) pour l'évaluation de notre proposition.

### *4.2.1.2. Interpolation hybride (INTER)*

Nous effectuons aussi nos études de l'adaptation non supervisée des modèles acoustiques en utilisant l'une des approches d'adaptation pour la parole non native appelée *interpolation*

*hybride* (interpolation et fusion) qui est proposée par [Tan, 2007]. On trouvera des informations plus détaillées sur cette approche dans la section 1.4.3.4 du chapitre 1.

Notre motivation pour l'étude de cette approche peut se résumer ainsi : au lieu d'utiliser une grande quantité de données, l'approche INTER génère un nouveau modèle acoustique multilingue adapté en utilisant un modèle acoustique de la langue cible (L2, la langue que le locuteur parle) et un modèle de la langue source (L1, la langue maternelle du locuteur). Par exemple, pour créer un modèle acoustique de la parole non native française parlée par les vietnamiens, [Tan, 2007] interpole le modèle acoustique de la langue cible française avec celui de la langue source vietnamienne.

À notre connaissance, les études sur l'adaptation des modèles acoustiques à base de l'approche d'interpolation des modèles sources et cibles précédentes [Tan, 2007; Wang, 2003] sont effectuées dans un contexte d'adaptation supervisée (les modèles source et cible sont connus à l'avance). D'ailleurs, selon les équations 1.9 à 1.11 du chapitre 1, l'interpolation hybride dépend nécessairement de trois paramètres : 1) le modèle acoustique de la langue cible ($MA_{cible}$) ; 2) le modèle acoustique de la langue source ($MA_{source}$) ; 3) le poids d'interpolation (w).

$$MA_{\text{inter}} = \{MA_{cible}, MA_{source}, w\} \tag{4.1}$$

Dans notre contexte d'adaptation non supervisée [Sam, 2010b], **ces trois paramètres ne sont pas connus à l'avance**. Par ailleurs, nous avons trois langues (le français, l'anglais et le vietnamien) à interpoler au lieu de deux langues seulement (une source et une cible) qui sont souvent étudiées dans les travaux précédents cités. Pour déterminer les trois paramètres de l'équation 4.1 à partir d'un segment de parole inconnue, nous proposons le processus suivant pour l'interpolation non supervisée des trois modèles acoustiques (EN, FR et VN).

- Définir les langues cibles et sources : la langue qui présente le meilleur score postérieur parmi les trois scores fournis par l'observateur de langue (OL) est considérée comme la langue cible (son modèle est $MA_{cible}$) ; les deux autres langues sont considérées comme les langues sources ($MA_{source1}$, $MA_{source2}$ pour les modèles acoustiques des langues qui ont, respectivement, un score postérieur au 2ème et 3ème rang) ;
- Définir le poids d'interpolation (w) : les scores postérieurs des langues sources sont considérés comme les poids d'interpolation. L'équation d'interpolation 4.1 devient alors :

$$MA_{\text{inter}} = \{MA_{cible}, MA_{source}, P(L_{source})\} \quad (4.2)$$

où $P(L_{source})$ est un score postérieur d'une langue parmi les deux langues sources ;

- Combiner les deux modèles interpolés : comme nous avons deux langues sources à interpoler, l'interpolation est faite deux fois successivement ; la première fois ($MA_{inter1}$) consiste à interpoler $MA_{cible}$ avec le modèle acoustique de la première langue source ($MA_{source1}$), et la deuxième fois ($AM_{inter2}$) interpole $MA_{cible}$ avec le modèle acoustique de la deuxième langue source ($MA_{source2}$) ; finalement ces deux modèles interpolés ($MA_{inter1}$ et $MA_{inter2}$) sont combinés, en utilisant la méthode de combinaison de modèles acoustiques ML-sep [Schultz, 2001], pour créer le modèle acoustique multilingue adapté ($MA\text{-}Mult_{adapté}$).

La figure 4.2 illustre le processus d'interpolation non supervisée dans le cas où notre observateur de langue (PR-VSM par exemple) donne les scores postérieurs P(EN) = 0,5 ; P(FR) = 0,3 ; P(VN)=0,2.

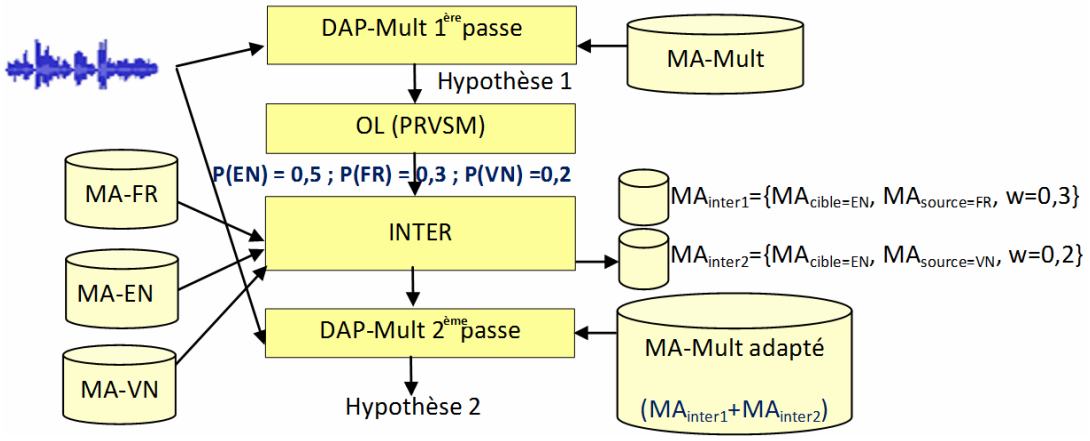
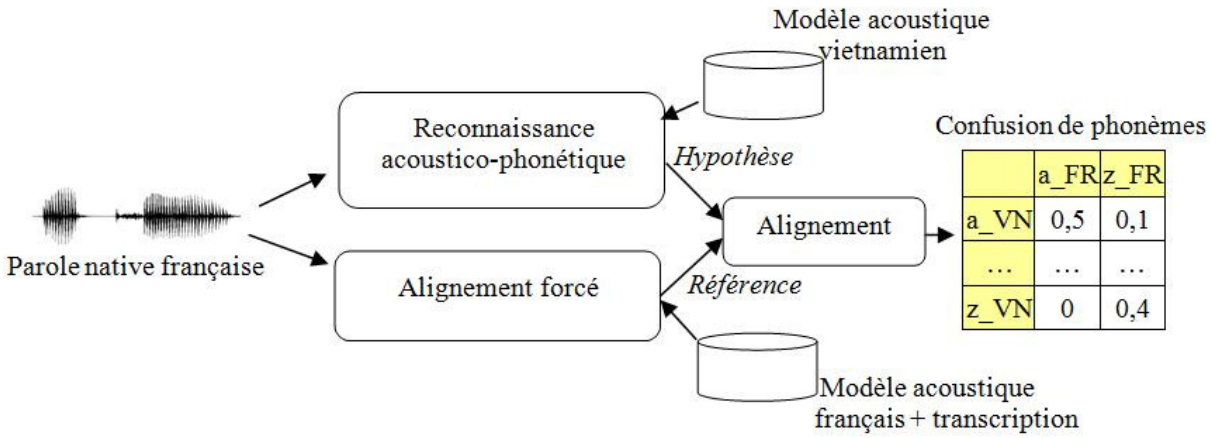



*Figure 4.2 : Exemple du processus d'interpolation non supervisée des trois modèles acoustiques*

Pour faire l'interpolation et la fusion de deux modèles acoustiques source et cible de manière supervisée, il est nécessaire d'avoir un tableau de substitution (ou tableau de correspondance) des phonèmes de la langue source vers la langue cible [Tan, 2008]. Dans notre tâche d'adaptation non supervisée de trois modèles acoustiques, il est obligatoire d'avoir six tableaux de substitution des phonèmes des langues sources vers les phonèmes des langues cibles (EN→FR, EN→VN, FR→EN, FR→VN, VN→EN, VN→FR). Ces tableaux peuvent être créés, soit à partir d'une connaissance linguistique fondée sur le tableau de l'Alphabet Phonétique International (API) [IPA, 1999], soit par la méthode automatique « matrice de confusion ».

Le processus de la matrice de confusion qui permet de construire le tableau de substitution des phonèmes sources en phonèmes cibles est le même que celui mentionné dans le chapitre 2 (figure 2.5) sauf que les données d'entrée et les modèles acoustiques utilisés sont différents. La figure 4.3 illustre la construction d'un tableau de substitution des phonèmes vietnamiens vers des phonèmes français selon la méthode automatique « matrice de confusion ».




| | a_FR | z_FR |
|---|---|---|
| a_VN | 0,5 | 0,1 |
| … | … | … |
| z_VN | 0 | 0,4 |

*Figure 4.3 : Exemple du transfert de phonèmes de la langue source VN à la langue cible FR (VN→FR) à base de la matrice de confusion*

Cependant, selon les résultats de substitution des phonèmes sur la matrice de confusion, les taux d'erreur de substitution sont assez grands (de 20 % à 30 %). Pour cette raison, nous avons décidé de créer les six tableaux de substitution majoritairement à partir du tableau API des trois langues traitées et de n'appliquer la méthode de la matrice de confusion que pour les phonèmes spécifiques des langues. Les tableaux 4.1a à 4.1f présentent les six tableaux de substitution des phonèmes des trois langues traitées.

Nous étudions également l'interpolation des modèles acoustiques suivie par l'approche MLLR (INTER-MLLR) en appliquant simplement une passe d'adaptation MLLR au modèle acoustique multilingue interpolé.

### 4.2.2. Adaptation « hors ligne » du modèle acoustique multilingue

#### *4.2.2.1. Same language Identification MLLR (SLI-MLLR)*

Il semble mieux de considérer plusieurs segments au lieu d'un seul segment de parole pour une adaptation MLLR. Ainsi, nous proposons un autre processus d'adaptation, appelé « MLLR par langue » ou *same language identification MLLR* (SLI-MLLR), qui utilise, dans son processus d'adaptation, non seulement le segment en cours de décodage, mais aussi les segments précédents (déjà décodés). L'objectif est de grouper les segments de même langue

parlée dans un groupe et de les utiliser dans le processus d'adaptation MLLR afin de mieux traiter le segment en cours de décodage.

Pour atteindre cet objectif, le processus d'adaptation se fait en trois étapes successivement :

- étape 1 : pour tous les segments déjà décodés, nous les regroupons en fonction des meilleurs scores postérieurs de langue fournis par le module « observateur » (donc on a au total trois groupes de langues : EN, FR et VN) ;
- étape 2 : en utilisant une décision de type « identification des langues » (LID ; la langue qui a le meilleur score est la langue parlée), si l'observateur indique que le segment de décodage en cours appartient à un des trois groupes de langues traitées (soit EN soit FR soit VN), tous les segments de parole et les hypothèses du décodeur première passe correspondant de ce groupe sont utilisés pour faire l'adaptation MLLR ;
- étape 3 : enfin, le segment en cours est décodé avec le modèle acoustique multilingue adapté.

Par exemple, si l'observateur du $n^{ème}$ segment donne P(EN) = 0,5 ; P(FR) = 0,3 et P(VN) = 0,2 ; alors SLI-MLLR indique que la langue parlée dans ce segment est l'anglais. Ainsi, il cherche dans tous les segments déjà décodés (du $1^{er}$ segment jusqu'à $(n-1)^{ème}$ segment) ceux qui ont P(EN) comme le meilleur score postérieur pour les utiliser dans l'adaptation MLLR.

#### *4.2.2.2. Phone mapping MLLR (PM-MLLR)*

La seule différence entre PM-MLLR et SL-MLLR est que PM-MLLR remplace, dans l'hypothèse utilisée pendant l'étape MLLR, tous les phonèmes des autres langues par les phonèmes similaires à la langue la plus probable selon l'observateur. La substitution de phonèmes (*phone mapping*) entre les trois langues est faite en se basant sur les six tableaux de correspondance mentionnés dans le tableau 4.1.

Par exemple, si l'hypothèse du décodeur, à la première passe sur un segment, est « h_EN e_EN l_FR o_EN », la langue détectée sur ce segment fournie par l'observateur est EN. Cela signifie que la technique PM-MLLR remplace le phonème /l_FR/ par le phonème similaire /l_EN/. Ensuite, PM-MLLR effectue le même processus d'adaptation en trois étapes que SLI-MLLR.

a.

| FR→EN | |
|---|---|
| a_FR | a_EN |
| aː_FR | aː_EN |
| ə_FR | ə_EN |
| b_FR | b_EN |
| d_FR | d_EN |
| e_FR | e_EN |
| ɛ_FR | ɛ_EN |
| f_FR | f_EN |
| g_FR | g_EN |
| h_FR | h_EN |
| i_FR | i_EN |
| j_FR | j_EN |
| k_FR | k_EN |
| l_FR | l_EN |
| m_FR | m_EN |
| n_FR | n_EN |
| ŋ_FR | ŋ_EN |
| ɔ_FR | ɔ_EN |
| o_FR | o_EN |
| p_FR | p_EN |
| s_FR | s_EN |
| ʃ_FR | ʃ_EN |
| SIL_FR | SIL_EN |
| t_FR | t_EN |
| u_FR | u_EN |
| v_FR | v_EN |
| w_FR | w_EN |
| z_FR | z_EN |
| ʒ_FR | ʒ_EN |
| ø_FR | u_EN |
| œ_FR | ɛ_EN |
| øœ_FR | ʊ_EN |
| ɑ_FR | ɛ_EN |
| ɑ̃_FR | o_EN |
| ɥ_FR | i_EN |
| ɛ̃_FR | ɛ_EN |
| ɲ_FR | ŋ_EN |
| ɔ̃_FR | o_EN |
| oɔ_FR | ʊ_EN |
| ʁ_FR | h_EN |
| œ̃_FR | æ_EN |
| y_FR | i_EN |
| ɛː_FR | ɛ_EN |

b.

| FR→VN | |
|---|---|
| a_FR | a_VN |
| b_FR | b_VN |
| d_FR | d_VN |
| e_FR | e_VN |
| ɛ_FR | ɛ_VN |
| f_FR | f_VN |
| h_FR | h_VN |
| i_FR | i_VN |
| j_FR | j_VN |
| ɲ_FR | ɲ_VN |
| k_FR | k_VN |
| l_FR | l_VN |
| m_FR | m_VN |
| n_FR | n_VN |
| ŋ_FR | ŋ_VN |
| ɔ_FR | ɔ_VN |
| o_FR | o_VN |
| p_FR | p_VN |
| s_FR | s_VN |
| SIL_FR | SIL_VN |
| t_FR | t_VN |
| u_FR | u_VN |
| v_FR | v_VN |
| w_FR | w_VN |
| z_FR | z_VN |
| ø_FR | ɤ_VN |
| œ_FR | ɤ_VN |
| øœ_FR | ɤ_VN |
| ɑ_FR | a_VN |
| aː_FR | a_VN |
| ə_FR | ɤ_VN |
| ɑ̃_FR | ɔ_VN |
| g_FR | ɣ_VN |
| ɥ_FR | w_VN |
| ɛ̃_FR | ɛχ_VN |
| ð_FR | ɔ_VN |
| oɔ_FR | o_VN |
| ʁ_FR | χ_VN |
| ʃ_FR | s_VN |
| œ̃_FR | a_VN |
| y_FR | u_VN |
| ʒ_FR | z_VN |
| ɛː_FR | ɛ_VN |

c.

| VN→EN | |
|---|---|
| a_VN | a_EN |
| b_VN | b_EN |
| d_VN | d_EN |
| e_VN | e_EN |
| ɛ_VN | ɛ_EN |
| f_VN | f_EN |
| h_VN | h_EN |
| i_VN | i_EN |
| j_VN | j_EN |
| k_VN | k_EN |
| l_VN | l_EN |
| m_VN | m_EN |
| n_VN | n_EN |
| ŋ_VN | ŋ_EN |
| ɔ_VN | ɔ_EN |
| o_VN | o_EN |
| p_VN | p_EN |
| s_VN | s_EN |
| SIL_VN | SIL_EN |
| t_VN | t_EN |
| u_VN | u_EN |
| v_VN | v_EN |
| w_VN | w_EN |
| z_VN | z_EN |
| ɤ_VN | ɛ_EN |
| ɤχ_VN | ɛ_EN |
| aχ_VN | a_EN |
| c_VN | ʃ_EN |
| cʰ_VN | tʃ_EN |
| ɛχ_VN | ɛ_EN |
| ɣ_VN | g_EN |
| ie_VN | ɛ_EN |
| ɲ_VN | ŋ_EN |
| ɯ_VN | u_EN |
| ɯɤ_VN | w_EN |
| ɔχ_VN | o_EN |
| ʂ_VN | ʃ_EN |
| tʰ_VN | f_EN |
| uo_VN | u_EN |
| χ_VN | h_EN |
| ʐ_VN | ʒ_EN |

d.

| VN→FR | |
|---|---|
| a_VN | a_FR |
| b_VN | b_FR |
| d_VN | d_FR |
| e_VN | e_FR |
| ɛ_VN | ɛ_FR |
| f_VN | f_FR |
| h_VN | h_FR |
| i_VN | i_FR |
| j_VN | j_FR |
| ɲ_VN | ɲ_FR |
| k_VN | k_FR |
| l_VN | l_FR |
| m_VN | m_FR |
| n_VN | n_FR |
| ŋ_VN | ŋ_FR |
| ɔ_VN | ɔ_FR |
| o_VN | o_FR |
| p_VN | p_FR |
| s_VN | s_FR |
| SIL_VN | SIL_FR |
| t_VN | t_FR |
| u_VN | u_FR |
| v_VN | v_FR |
| w_VN | w_FR |
| z_VN | z_FR |
| ɤ_VN | œ_FR |
| ɤχ_VN | œ_FR |
| aχ_VN | a_FR |
| c_VN | ʃ_FR |
| cʰ_VN | ʃ_FR |
| ɛχ_VN | ɛ̃_FR |
| ɣ_VN | g_FR |
| ie_VN | ɛː_FR |
| ɯ_VN | u_FR |
| ɯɤ_VN | w_FR |
| ɔχ_VN | o_FR |
| ʂ_VN | ʃ_FR |
| tʰ_VN | f_FR |
| uo_VN | u_FR |
| χ_VN | ʁ_FR |
| ʐ_VN | z_FR |

e.

| EN→FR | |
|---|---|
| ə_EN | ə_FR |
| a_EN | a_FR |
| aː_EN | aː_FR |
| b_EN | b_FR |
| d_EN | d_FR |
| e_EN | e_FR |
| ɛ_EN | ɛ_FR |
| f_EN | f_FR |
| g_EN | g_FR |
| h_EN | h_FR |
| i_EN | i_FR |
| j_EN | j_FR |
| k_EN | k_FR |
| l_EN | l_FR |
| m_EN | m_FR |
| n_EN | n_FR |
| ŋ_EN | ŋ_FR |
| ɔ_EN | ɔ_FR |
| o_EN | o_FR |
| p_EN | p_FR |
| s_EN | s_FR |
| ʃ_EN | ʃ_FR |
| t_EN | t_FR |
| u_EN | u_FR |
| v_EN | v_FR |
| w_EN | w_FR |
| z_EN | z_FR |
| ʒ_EN | ʒ_FR |
| SIL_EN | SIL_FR |
| æ_EN | œ̃_FR |
| aɪ_EN | aː_FR |
| ð_EN | d_FR |
| dʒ_EN | ʒ_FR |
| ɪ_EN | ɛ_FR |
| ɔɪ_EN | ɔ̃_FR |
| r_EN | ɥ_FR |
| θ_EN | t_FR |
| tʃ_EN | ʃ_FR |
| ʊ_EN | øœ_FR |
| ʌ_EN | oɔ_FR |

f.

| EN→VN | |
|---|---|
| a_EN | a_VN |
| b_EN | b_VN |
| d_EN | d_VN |
| e_EN | e_VN |
| ɛ_EN | ɛ_VN |
| f_EN | f_VN |
| h_EN | h_VN |
| i_EN | i_VN |
| j_EN | j_VN |
| k_EN | k_VN |
| l_EN | l_VN |
| m_EN | m_VN |
| n_EN | n_VN |
| ŋ_EN | ŋ_VN |
| ɔ_EN | ɔ_VN |
| o_EN | o_VN |
| p_EN | p_VN |
| s_EN | s_VN |
| t_EN | t_VN |
| u_EN | u_VN |
| v_EN | v_VN |
| w_EN | w_VN |
| z_EN | z_VN |
| SIL_EN | SIL_VN |
| æ_EN | a_VN |
| ə_EN | ɤ_VN |
| aː_EN | ɔ_VN |
| aɪ_EN | a_VN |
| ð_EN | d_VN |
| dʒ_EN | z_VN |
| g_EN | ɣ_VN |
| ɪ_EN | e_VN |
| ɔɪ_EN | ɔ_VN |
| r_EN | w_VN |
| ʃ_EN | s_VN |
| θ_EN | t_VN |
| tʃ_EN | cʰ_VN |
| ʊ_EN | ɤ_VN |
| ʌ_EN | o_VN |
| ʒ_EN | ʐ_VN |

*Tableau 4.1 : Six tableaux de substitution des phonèmes (les couleurs grises représentent les transferts de phonèmes de la langue source à la langue cible en se basant sur la méthode par matrice de confusion, le reste étant obtenu selon l'API)*

## 4.3. Méthodes d'évaluation de la performance des systèmes

### 4.3.1. Représentation phonétique

La représentation phonétique est importante dans notre contexte d'étude, car les évaluations des performances des systèmes d'adaptation sont faites au niveau du phonème (les hypothèses des décodeurs acoustico-phonétiques sont les séquences de phonèmes).

La représentation phonétique des langues peut être exprimée à l'aide de l'Alphabet Phonétique International (API) ou d'autres représentations phonétiques standard comme X-SAMPA [Wells, 1995] et Worldbet [Hieronymus, 1993]. Les deux dernières encodent les symboles API en caractèreS ASCII[12]. Avec la présentation des symboles en ASCII, X-SAMPA et Worldbet assurent la lisibilité des caractères API dans les outils d'encodage textuels qui ne sont pas compatibles avec les caractères Unicode [Aliprand, 2003]. Les symboles API sont encodés par les caractères Unicode. La représentation phonétique X-SAMPA est utilisée pour encoder les phonèmes dans notre contexte d'étude de l'adaptation des modèles acoustiques multilingues, car l'outil sphinx3 ne supporte pas les caractères Unicode, surtout dans leur modèle lexical (le dictionnaire de prononciation). Le lecteur trouvera le tableau de la présentation phonétique X-SAMPA correspondant au tableau API dans l'annexe B.

Il est aussi important de préciser que, dans le processus de décodage, un reconnaisseur acoustico-phonétique multilingue produit ses résultats (hypothèses) sous forme de séquences de phonèmes étiquetés par les langues auxquelles ces phonèmes appartiennent (par exemple, le phonème /a/ de la langue anglaise (EN) est représenté par /a_EN/). Dans le processus d'évaluation de la performance du système, pour réduire la disparité (*mismatch*) entre les phonèmes communs aux langues traitées, les étiquettes qui indiquent les langues d'origine des phonèmes sont supprimées. En effet, le but est d'évaluer la capacité du décodeur acoustico-phonétique multilingue à générer une suite de phonèmes correcte. Dans le dernier cas, les phonèmes /a_EN/, /a_FR/ et /a_VN/ dans des séquences phonétiques de la référence ou de l'hypothèse sont représentés par un seul phonème /a/ par exemple.

### 4.3.2. Métrique d'évaluation

Dans cette section, nous présentons plus en détail l'outil que nous utilisons pour évaluer la performance des décodeurs acoustico-phonétiques multilingues. La plate-forme d'évaluation

[12] http://en. wikipedia. org/wiki/ASCII

des outils de reconnaissance de la parole SCTK inclut l'outil « *sclite* » qui implémente un algorithme de programmation dynamique pour calculer un taux d'erreur de phonèmes dans le meilleur des cas entre une phrase de référence et la phrase correspondante (hypothèse), en prenant en compte les insertions, omissions et substitutions. Le processus d'évaluation de l'outil « sclite » est en deux étapes :

- étape 1 - alignement textuel : « sclite » utilise un algorithme de programmation dynamique pour minimiser la distance de Levenshtein entre deux chaînes de texte (une référence et son hypothèse correspondante). Le lecteur trouvera le détail de l'algorithme d'alignement dynamique dans [Sankoff, 1999] ;
- étape 2 - *scoring* : après avoir aligné les chaînes de référence et d'hypothèse, les taux d'erreur des phonèmes (*Phone Error Rate*, PER) sont calculés selon l'équation suivante :

$$PER = \frac{I + O + S}{N} * 100 \tag{4.3}$$

où I, O, S représentent respectivement l'insertion, l'omission et la substitution ; N est le nombre total de phonèmes dans la référence.

Par exemple, si la séquence phonétique de référence du segment de la parole « *ok, c'est bien !* » est « *o k e s EE b j in* » et si celle de l'hypothèse correspondante est « *o k e e s E j i N* » (ces deux séquences phonétiques étant représentées ici dans le standard X-SAMPA) ; alors, l'outil « sclite » évalue la performance du décodeur acoustico-phonétique comme illustré dans la figure 4.4 ci-dessous.

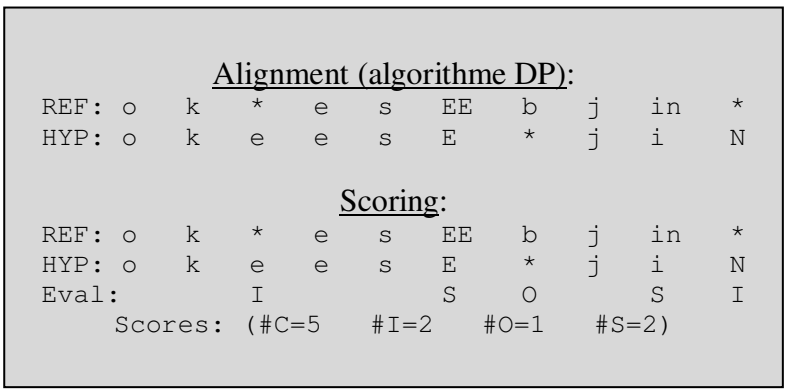



*Figure 4.4 : Exemple d'évaluation de la performance du décodeur acoustico-phonétique selon l'outil « sclite » (le symbole «* » représente une erreur, #C, #I, #O et #S sont le nombre de mots corrects, insertions, omissions et substitutions des phonèmes entre la référence et l'hypothèse)*

Selon la figure 4.4, le taux d'erreur des phonèmes, dans ce cas, est égal au rapport entre le nombre des phonèmes erronés (#I + #O + #S) et le nombre total des phonèmes dans la référence. Cela rend le taux d'erreur égal à 62,5 % (PER = (5/8)*100).

## 4.4. Résultats d'expérimentation

### 4.4.1. Système acoustico-phonétique multilingue de référence (*Baseline*)

Nous présentons, dans cette section, le système de référence (*baseline*) pour lequel nous allons comparer la performance avec celles des systèmes adaptés (les décodeurs qui utilisent les modèles acoustiques multilingues adaptés). Nous rappelons que l'évaluation de la performance des systèmes acoustico-phonétiques est fondée sur les taux d'erreur de phonèmes fournis par l'outil sclite (section 4.3.2). Dans notre contexte d'étude, le système de référence est un décodeur acoustico-phonétique multilingue simple (le décodeur acoustico-phonétique multilingue illustré dans la figure 3.3 du chapitre 3).

Avant de comparer les résultats des systèmes adaptés avec ceux du système acoustico-phonétique multilingue de référence (DAP-Mult), nous comparons, tout d'abord, les taux d'erreur de phonèmes du DAP-Mult avec ceux des trois systèmes acoustico-phonétiques monolingues (celui de l'anglais (DAP-EN), du français (DAP-FR) et du vietnamien (DAP-VN)). La comparaison entre les résultats des systèmes acoustico-phonétiques monolingues et ceux du système multilingue est faite de deux façons différentes.

- Décodage non supervisé des segments de parole. Nous utilisons, dans ce cas, des systèmes acoustico-phonétiques monolingues pour décoder tous les segments de la parole de test sans savoir à l'avance les langues parlées. Cela signifie que chaque reconnaisseur monolingue décode toute la parole native et non native anglaise, française et vietnamienne disponibles dans le corpus de test. Le tableau 4.2 présente la comparaison des résultats des reconnaisseurs monolingues et multilingues de référence (*baseline*) en terme du décodage non supervisé des segments de la parole ;
- Décodage supervisé des segments de la parole : dans ce cas, les segments de la parole native et non native sont groupés par langue parlée avant que la procédure de décodage monolingue ne soit mise en œuvre. Par exemple, les segments de parole native et non native de l'anglais (ENen, ENfr et ENvn) sont décodés par le système DAP-EN. Les résultats du décodage supervisé sont résumés dans le tableau 4.3.

| | **DAP-Mult (*baseline*)** | **DAP-EN** | **DAP-FR** | **DAP-VN** |
|---|---|---|---|---|
| **ENen** | 57,5 | **51** | 67,9 | 74,7 |
| **ENfr** | 60,2 | **55,9** | 63,1 | 65,8 |
| **ENvn** | 57,6 | **52,7** | 63,2 | 65,1 |
| **FRfr** | 56,7 | 67,9 | **52,2** | 69,6 |
| **FRen** | 59,1 | 71,9 | **56,1** | 76,9 |
| **FRvn** | 55,8 | 62,6 | **54,6** | 61,8 |
| **VNvn** | 47,3 | 68,7 | 68,8 | **44,7** |
| **VNen** | 57 | 63 | 68,3 | **51,3** |
| **VNfr** | 48,9 | 65 | 64 | **46,9** |
| **Moyenne** | **55,8** | 61,6 | 61,5 | 61,9 |

*Tableau 4.2 : Comparaison des PERs (%) des reconnaisseurs en vue du décodage non supervisé des segments de parole (les chiffres en gras représentent le meilleur score)*

| | **DAP-Mult (*baseline*)** | **DAP-Mono (*supervisé*)** |
|---|---|---|
| **ENen** | 57,5 | 51 |
| **ENfr** | 60,2 | 55,9 |
| **ENvn** | 57,6 | 52,7 |
| **FRfr** | 56,7 | 52,2 |
| **FRen** | 59,1 | 56,1 |
| **FRvn** | 55,8 | 54,6 |
| **VNvn** | 47,3 | 44,7 |
| **VNen** | 57 | 51,3 |
| **VNfr** | 48,9 | 46,9 |
| **Moyenne** | 55,8 | 51,5 |

*Tableau 4.3 : Comparaison des PERs (%) du décodeur acoustico-phonétique DAP-Mult et des DAP monolingues (dans le cas où l'on suppose connue la langue parlée, i.e. décodage supervisé des segments de la parole)*

Il est important de préciser qu'il y a 69 % des données de test qui sont de la parole non native. De plus, nous n'utilisons que des systèmes acoustico-phonétiques (les systèmes qui ne dépendent que des modèles acoustiques) dans nos processus de décodage des segments de la parole, ce qui explique pourquoi les taux d'erreur de phonèmes sont grands (environ 50 %) pour les décodeurs acoustico-phonétiques monolingues ainsi que pour le décodeur multilingue « baseline ».

Dans le contexte du décodage non supervisé (les langues parlées ne sont pas connues à l'avance ; voit tableau 4.2), nous observons que les décodeurs monolingues donnent les meilleurs résultats pour leurs segments de parole native et non native, mais leurs taux d'erreur de phonèmes (PER) sont plus élevés pour les segments contenant de la parole d'une autre langue. Par exemple, le décodeur acoustico-phonétique de l'anglais (DAP-EN) fournit des taux d'erreur de phonèmes (PER) de parole native et non native en langue anglaise (ENen, ENfr et ENvn) inférieurs à 56 %, tandis que les taux d'erreur sur de la parole française et vietnamienne sont supérieurs à 62 %. Cela est bien évidemment dû au fait que tous les phonèmes français et vietnamiens ne peuvent être reconnus par un système monolingue anglais. La moyenne des taux d'erreur fournis par le système de référence (DAP-Mult), au contraire, est significativement plus faible que celle fournie par les décodeurs monolingues. Cela met en évidence le fait que le décodeur multilingue (DAP-Mult) donne la meilleure performance de décodage phonétique par rapport aux décodeurs monolingues, pour les segments de parole dont la langue parlée est inconnue.

Quand les langues des segments de parole sont connues à l'avance (décodage supervisé), le système multilingue dégrade un petit peu la performance du système monolingue (environ 4 % de différence). Cela peut être causé par le faite que le modèle acoustique monolingue ne contient que les phonèmes plus spécifiques à la langue traitée, tandis que le modèle multilingue contient plus de phonèmes que nécessaires, ce qui rend les confusions plus nombreuses.

D'après le tableau 4.2 et le tableau 4.3, nous pouvons conclure que les décodeurs acoustico-phonétiques multilingues semblent meilleurs que les décodeurs monolingues dans notre contexte d'étude de l'adaptation des modèles acoustiques pour laquelle les langues des segments à décoder sont inconnus. Cela aussi met en évidence que notre choix d'utiliser le modèle acoustique multilingue au lieu du modèle acoustique monolingue dans notre processus d'adaptation autonome est correct.

### 4.4.2. Adaptation du modèle acoustique multilingue (MA-Mult)

Nous évaluons tout d'abord les résultats des décodeurs qui utilisent les scores postérieurs fournis par l'observateur de langue PR-VSM (section 3.4 du chapitre 3) pour aider le processus d'adaptation non supervisée. Le tableau 4.4 résume la comparaison des taux d'erreur de phonèmes (PER) du décodeur de référence et ceux des systèmes adaptés en utilisant l'observateur de langues PR-VSM.

Afin d'évaluer plus profondément les adaptations (non supervisées) utilisant l'observateur de langues, nous étudions également la performance d'adaptation en utilisant une identification parfaite de la langue parlée (par « oracle ») dans le module de l'observateur de langue. Dans ce cas, les adaptations non supervisées sont toujours établies en utilisant les scores postérieurs générés par l'observateur de langues PR-VSM, mais la langue parlée du segment (langue cible, L2) est considérée comme connue à l'avance (la langue cible détectée par PR-VSM n'est pas considérée).

| | **Référence** | **Adaptation « en ligne »** | | | **Adaptation « hors ligne »** | |
|---|---|---|---|---|---|---|
| | **(*Baseline*)** | **MLLR** | **INTER** | **INTER-MLLR** | **SLI-MLLR** | **PM-MLLR** |
| **ENen** | 57,5 | 56,2 | 59,2 | 59,2 | 56,9 | **54,7** |
| **ENfr** | 60,2 | 60,4 | **54,3** | **54,3** | 58,9 | 55,4 |
| **ENvn** | 57,6 | 57,6 | **52,7** | 52,8 | 57,0 | 56,0 |
| **FRfr** | 56,7 | 56,5 | 58,2 | 57,8 | 55,3 | **54,8** |
| **FRen** | 59,1 | 58,5 | 51,7 | **51,3** | 57,1 | 55,2 |
| **FRvn** | 55,8 | 55,9 | **55,0** | **55,0** | 55,4 | **55,0** |
| **VNvn** | **47,3** | 45,7 | 53,2 | 53,2 | 47,5 | 47,6 |
| **VNen** | 57,0 | 56,5 | 53,4 | **53,2** | 57,3 | 57,2 |
| **VNfr** | 48,9 | 49,0 | 45,9 | **45,8** | 51,5 | 51,2 |
| **Moyenne** | 55,8 | 55,4 | 53,3 | **53,2** | 55,5 | 54,3 |

*Tableau 4.4 : PER (%) des différents systèmes acoustico-phonétiques de référence et adaptés en utilisant l'observateur de langue PR-VSM (les colonnes en gris représentent les résultats de référence des décodeurs indépendants de l'observateur de langues, et les chiffres en gras de chaque ligne représentent les meilleurs scores)*

Par exemple, supposons que le segment à décoder contienne la langue anglaise parlée par un locuteur français, et que PR-VSM produise les scores postérieurs P(FR) = 0,5 ; P (EN) = 0,4 et P (VN) = 0,1. Dans le cas *oracle*, bien que le score postérieur de la langue anglaise P(EN) ne soit pas le meilleur score, l'anglais est tout de même considéré comme la langue cible avec un poids de 0,4, tandis que les autres langues (FR, VN) sont considérées comme des langues source (L1) avec les poids 0,5 et 0,1 respectivement. Il est nécessaire de préciser que ce qui est important dans le processus d'interpolation est de définir la langue cible, car nous devons substituer (*phone mapping*) tous les phonèmes des langues sources en phonèmes de la langue cible. Dans ce dernier exemple, l'observateur de langue (OL) fournit la langue cible incorrectement.

Le tableau 4.5 résume la performance des systèmes adaptés pour cette condition « oracle » de l'observateur de langues.

| | **Référence** | **Adaptation « en ligne »** | | | **Adaptation « hors ligne »** | |
|---|---|---|---|---|---|---|
| | **(*Baseline*)** | **MLLR** | **INTER** | **INTER-MLLR** | **SLI-MLLR** | **PM-MLLR** |
| **ENen** | 57,5 | 56,2 | 59,5 | 59,2 | 56,0 | **53,9** |
| **ENfr** | 60,2 | 60,4 | 52,9 | **52,4** | 57,2 | 55,1 |
| **ENvn** | 57,6 | 57,6 | 50,0 | **49,6** | 56,8 | 55,8 |
| **FRfr** | 56,7 | 56,5 | 58,2 | 57,8 | 55,3 | **54,8** |
| **FRen** | 59,1 | 58,5 | 50,5 | **49,9** | 57,1 | 55,2 |
| **FRvn** | 55,8 | 55,9 | 49,7 | **48,8** | 53,5 | 53,3 |
| **VNvn** | 47,3 | 45,7 | 51,9 | 50,4 | **46,8** | **46,8** |
| **VNen** | 57,0 | 56,5 | 48,6 | **48,3** | 56,0 | 53,9 |
| **VNfr** | 48,9 | 49,0 | 43,9 | **43,7** | 46,7 | 46,3 |
| **Moyenne** | 55,8 | 55,4 | 51,1 | **50,5** | 54,1 | 52,9 |

*Tableau 4.5 : PER (%) des différences systèmes acoustico-phonétiques de référence et adaptés en supposant parfait l'observateur de langues (les colonnes en gris représentent les résultats des décodeurs indépendants de l'observateur de langue et les chiffres en gras de chaque ligne représentent les meilleurs scores)*

Grâce aux performances des adaptations présentées dans les tableaux 4.3 et 4.4, nous faisons les observations suivantes.

- L'adaptation « en ligne » MLLR diminue peu les taux d'erreur des phonèmes du système de référence car elle utilise un seul segment (le segment en cours de décodage) dans son processus d'adaptation.
- Les adaptations « hors ligne » SLI-MLLR et PM-MLLR donnent des résultats très positifs si la performance de l'observateur PR-VSM est bonne, par exemple sur la parole native : l'adaptation PM-MLLR diminue les PERs de ENen et FRfr d'environ 2 % à 3 % en les comparant avec ceux du système de référence. Cependant, cela n'est pas vrai pour le vietnamien natif (VNvn). Cela s'explique par le fait que, si nous observons la performance de l'observateur de langue PR-VSM, celui-ci ne donne pas la bonne décision pour les segments VNvn. Par ailleurs, l'adaptation PM-MLLR dégrade les PERs des segments VNvn du système de référence dans le cas de l'observateur parfait.
- Dans le cas des adaptations « en ligne » fondées sur l'observateur de langues PR-VSM (tableau 4.4), les adaptations INTER et INTER-MLLR diminuent les PER du système de référence pour chaque parole non native, malgré le fait que la performance de l'observateur de langues ne soit pas bonne dans certain cas (les segments VNen et VNfr par exemple) ;
- Dans le cas parfait de l'observateur de langue, l'adaptation « en ligne » INTER et INTER-MLLR diminue significativement, pour chaque parole non native, non seulement les PER du système de référence (tableau 4.5), mais aussi les PER des décodeurs monolingues (DAP-EN, DAP-FR et DAP-VN), pour lesquels une parfaite identification de la langue parlée est supposée pour tous les segments avant de les décoder (figure 4.5).
- Pour la parole non native, même si les taux d'erreur de phonèmes de FRvn et VNen sont plus élevés que ceux des décodeurs monolingues (à cause de la performance limitée de PR-VSM pour ces groupes de parole), la moyenne des taux d'erreur de l'adaptation INTER-MLLR fondée sur l'observateur PR-VSM est plus faible que celle des décodeurs monolingues (DAP-EN, DAP-FR et DAP-VN), dans lesquels une parfaite identification de la langue parlée est considérée pour tous les segments avant de les décoder (figure 4.5). Cela confirme que l'adaptation non supervisée INTER-MLLR fondée sur l'observateur PR-VSM que nous avons proposée est une alternative intéressante pour décoder la parole non native.

- Au contraire, l'adaptation « en ligne » INTER et INTER-MLLR dégrade généralement les performances du système de référence si les segments décodés contiennent de la parole native (tableau 4.4 et 4.5).

Avec ces observations, nous pouvons conclure que l'adaptation autonome d'un modèle multilingue INTER-MLLR que nous avons proposée est un choix prometteur pour décoder La parole non native de plusieurs langues, pour différentes origines de locuteurs, malgré une performance de l'observateur de langue non optimale. Malheureusement, cette technique dégrade la performance du système de référence si les segments à décoder sont de la parole native. Cela nous amène à envisager un processus de discrimination entre parole native et parole non native avant d'utiliser les techniques d'adaptation autonome. Dans ce scénario, la parole détectée comme native serait décodée avec les décodeurs monolingues correspondants, et la parole détectée comme non native serait décodée avec les systèmes adaptés en utilisant l'adaptation INTER ou INTER-MLLR. Le chapitre suivant (chapitre 5) présente une étude préliminaire sur les différentes techniques de détection de parole native et non native.

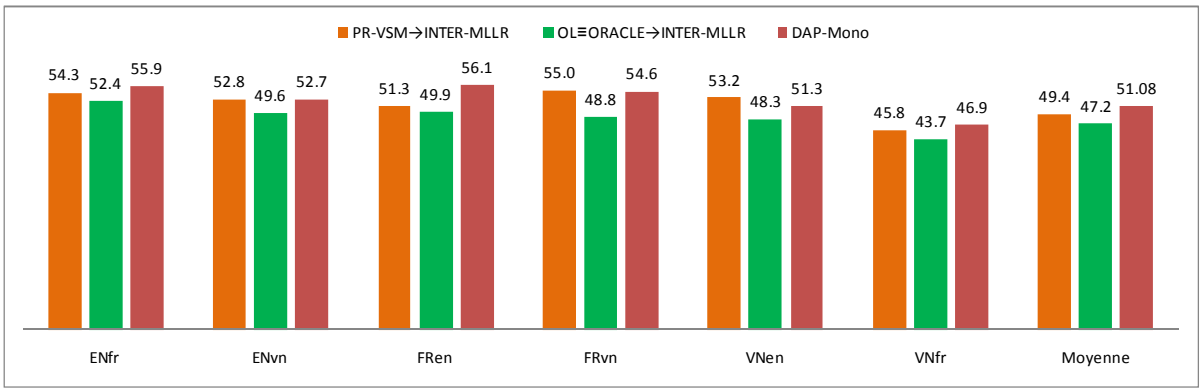



*Figure 4.5 : Comparaison des PERs (%) de INTER-MLLR fondée sur l'observateur PR-VSM et des DAP-Mono (ces derniers utilisent une identification parfaite des langues parlées dans les segments décodés)*

## 4.5. Conclusion

Dans ce chapitre, nous avons présenté différentes techniques d'adaptation des modèles acoustiques multilingues dans un contexte où les données d'adaptation sont indisponibles et les langues parlées (ainsi que l'origine des locuteurs) des segments à décoder sont inconnues.

Selon les résultats d'expérimentation, les techniques d'interpolation des modèles acoustiques (INTER et INTER-MLLR) que nous avons proposées, fondées sur l'observateur de langues PR-VSM, sont des choix prometteurs pour traiter la parole non native ; elles dégradent, en revanche, les performances sur de la parole native.

Pour rendre le processus de décodage phonétique plus robuste pour les deux types de la parole, native et non native, nous pensons qu'un module de discrimination entre parole native et parole non native peut être utile en début de chaîne. Nous présentons nos études préliminaires sur les différentes techniques de discrimination possible dans le chapitre suivant.

# Chapitre 5 :
# Premières études sur la discrimination entre parole native et non native

## 5.1. Introduction

Selon les conclusions de nos chapitres précédents, l'adaptation autonome des modèles acoustiques multilingues comme INTER-MLLR est un choix prometteur pour décoder la parole non native de plusieurs langues parlées par des locuteurs de différentes origines, sans besoin de données d'adaptation. Malheureusement, cette technique dégrade aussi la performance du système de référence si les segments à décoder sont de la parole native. Afin de rendre les systèmes adaptés plus robustes pour tous les groupes de parole (native et non native), nous proposons un module de discrimination entre parole native et non native pour générer deux groupes : le groupe de parole native sera décodé par les décodeurs acoustico-phonétiques monolingues (DAP-Mono) correspondants, tandis que le groupe de parole non native sera décodé en utilisant les systèmes adaptés selon notre technique d'adaptation autonome.

Dans nos études préliminaires de discrimination entre parole native et non native [Sam, 2011] , nous étudions différentes techniques de discrimination y compris les techniques se basant sur les caractéristiques phonétiques du signal (pitch et formants) qui sont robustes (selon [Arslan, 1996]) aux différences de canal entre les données d'entraînement et les données de test. Construire un tel système de détection nécessite un corpus, or seules des données de parole native et non native française [Tan, 2006] sont disponibles. Nous effectuons donc nos premières études seulement pour de la parole native et non native française. Dans ce chapitre, les corpus de test et d'entraînement sont détaillés dans la section 5.2. Les techniques de discrimination entre parole native et non native française et les résultats d'expérimentation sont présentés dans la section 5.3 et la section 5.4 respectivement. L'utilisation de ce module de détection native et non native en amont du processus d'adaptation autonome est présentée à la fin du chapitre (section 5.5). Enfin, il est important de noter que le travail présenté dans ce chapitre a été fait en collaboration avec un post-doctorant de la NTU de Singapour, visiteur au LIG sur une période de trois mois.

# 5.2. Corpus

### 5.2.1. Choix des bases de données

| **Origine** | Français | Vietnamien | Chinois | Anglais | Cambodgien | Total par BD |
|---|---|---|---|---|---|---|
| **BD1** | *2695* (3) | *10203* (8) | *11296* (8) | *0* (0) | *0* (0) | *24194* (19) |
| **BD2** | *6600* (12) | *0* (0) | *0* (0) | *0* (0) | *0* (0) | *6600* (12) |
| **BD3** | *2797* (3) | *1177* (4) | 0 (0) | *584* (1) | *1822* (1) | *6380* (9) |
| **Total par origine** | *12092* (18) | *11380* (12) | *11296* (8) | *584* (1) | *1822* (1) | *37174* (40) |

*Tableau 5.1 : Quantité de données (les chiffres en italique sont les valeurs en secondes) et les nombres de locuteurs (les chiffres entre parenthèses) des trois bases de données utilisées*

Dans nos études de discrimination entre parole native et non native française, nous extrayons les signaux de parole à partir de trois corpus différents.

- Première base de données (BD1) : nous prenons toutes les données du corpus de parole native et non native française créé au sein de l'équipe (GETALP) du laboratoire d'informatique de Grenoble (LIG) dans le cadre de la thèse de Tien Ping Tan [Tan, 2008]. Ce corpus contient les enregistrements de trois locuteurs natifs et 16 locuteurs non natifs (huit vietnamiens et huit chinois) qui correspondent à environ 6 h 40 mn de signal au total ; plus de détails peuvent être trouvés dans [Tan, 2006].
- Deuxième base de données (BD2) : comme BD1 ne contient que trois locuteurs natifs qui ne sont pas suffisants pour entraîner les systèmes de discrimination entre parole native et non native française, nous extrayons les locuteurs natifs à partir d'un corpus enregistré pour le développement d'un « livre de phrases » français / khmer pour smartphones [Touch, 2010]. La BD2 contient 12 locuteurs natifs français correspondant à une durée totale de 1 h 50 mn de signal.
- Troisième base de données (BD3) : nous extrayons aussi le signal de parole native et non native en français à partir du corpus MICA-MultiMeet mentionné dans le chapitre

2 de ce mémoire. La BD3 contient trois locuteurs natifs et six non natifs (quatre Vietnamiens, un Anglais et un Cambodgien).

Le tableau 5.1 présente les quantités de parole native et non native des trois bases de données que nous allons utiliser dans notre expérimentation (section 5.4).

Les trois bases de données contiennent toutes le même type de parole (parole lue). Par contre, le défi ici est la distorsion de canal entre deux des trois bases de données. Nous discuterons plus en détail ce sujet dans les sections suivantes.

### 5.2.2. Protocoles

Avec ces trois bases de données, nous définissons deux protocoles différents (tâches T1 et T2). Dans la tâche 1 (T1), nous utilisons les BD1 et BD2 pour entraîner le modèle et aussi pour le tester. Il y a donc au total 15 locuteurs natifs et 16 locuteurs non natifs français dans BD1 et BD2. Pour produire des résultats solides, nous avons utilisé la méthode de validation croisée (*cross validation*) avec 15 configurations consécutives. Pour chacune des 14 premières configurations, la parole d'un locuteur natif et celle d'un locuteur non natif sont utilisées pour tester les modèles et les enregistrements des autres locuteurs sont utilisés pour entraîner les modèles. Dans la dernière configuration (la 15$^{e}$), les enregistrements d'un locuteur natif et de deux locuteurs non natifs sont utilisés pour tester les modèles. Ce protocole permet de tester tous les locuteurs avec les modèles entraînés par les autres locuteurs.

Un des facteurs importants que nous allons étudier est la différence entre les conditions d'enregistrement sur l'ensemble de données. Comme la majorité des locuteurs natifs utilisés (12 locuteurs qui sont dans le BD2) sont enregistrés dans des conditions différentes que pour les signaux de BD1 qui contient tous les locuteurs non natifs et très peu de locuteurs natifs (3 locuteurs seulement), il y a un danger que le modèle apprenne à classer des conditions d'enregistrement plutôt qu'à discriminer entre parole native et parole non native.

Pour évaluer la robustesse des différentes approches proposées, à travers des conditions d'enregistrement différentes, nous définissons une autre tâche (T2), dans laquelle nous testons les systèmes de détection qui sont entraînés par les données extraites de BD1 et BD2 sur la base BD3. Ainsi, si un système détecte les conditions d'enregistrement plutôt que la parole non native, il devrait échouer à la tâche 2 car les conditions d'enregistrement de BD3 sont différentes de celles des BD1 et BD2. Notre but est alors, de chercher des approches de discrimination robustes et efficaces pour les deux tâches définies ci-dessus.

## 5.3. Approches de discrimination entre parole native et non native

Dans cette section, nous présentons nos études sur les différentes approches à la discrimination entre parole native et non native. Nous étudions cinq approches à la discrimination qui correspondent aux trois types d'analyse des caractéristiques de la parole : 1) les approches fondées sur les vecteurs cepstraux comme les coefficients MFCC (*Mel Frequency-warped Cepstral Coefficients*) [Huang, 2001] et la modulation du spectre proposée par [Kinnunen, 2006] ; 2) les approches à base de caractéristiques prosodiques comme le pitch (la fréquence fondamentale) et formantiques comme les valeurs des deux premiers formants, qui sont largement utilisées dans les études de détection d'accent des locuteurs étrangers [Grover, 1987; Hansen, 1995; Zissman, 1996b] ; 3) l'approche phonotactique PR-VSM mentionnée dans le chapitre 3.

Il est nécessaire de préciser que notre objectif est de décider si un segment du signal en entrée est de la parole native ou non native. Les approches étudiées vont donc analyser entièrement les caractéristiques des pitch, des formants et des cepstres du segment de parole au lieu des caractéristiques de chaque phonème du segment.

### 5.3.1. MFCC

Nous étudierons des caractéristiques des signaux fondées sur les coefficients cepstraux MFCC dans notre processus de discrimination entre parole native et non native. Dans ce cas, pour capturer les changements temporels (les caractéristiques dynamiques) dans le spectre d'un segment de parole, les première et deuxième dérivées des vecteurs caractéristiques MFCCs sont calculées et sont concaténées avec les MFCC pour produire un vecteur caractéristique final de dimension 3×M. En général, les systèmes de reconnaissance automatique de la parole utilisent 13 coefficients MFCC (M = 13) [Huang, 2001]. Un cas particulier est [Piat, 2008], qui a utilisé 12 coefficients MFCC (le paramètre « énergie » n'est pas considéré) dans son travail de discrimination entre parole non native. Par conséquent, après concaténation avec les coefficients de delta et d'accélération, le vecteur caractéristique final de MFCC présente une dimension de 36.

Comme les coefficients MFCC sont très influencés par les distorsions de bruit et de canal, nous appliquons la technique d'égalisation d'histogrammes (*histogram equalization*, HEQ), qui a été initialement détaillée et utilisée en traitement des images [Gonzalez, 2002], à toutes

les données d'entraînement et de test pour réduire les effets causés par la distorsion de canal entre les données d'entraînement et de test.

### 5.3.2. Modulation du spectre

Nous étudions aussi une des approches basées sur l'analyse spectrale du signal dans notre tâche de discrimination. Cette approche s'appelle « la modulation spectrale » et consiste à étudier les informations temporelles à long terme du signal de parole. Elle a été utilisée dans un système de reconnaissance du locuteur [Kinnunen, 2006].

La génération des paramètres de modulation spectrale est illustrée par la figure 5.1. La partie gauche de la figure présente les coefficients d'un groupe de filtres Mel (*Mel Filterbank*) d'un segment de parole. Si nous traitons la trajectoire de chaque banc de filtres (*Filterbank*) comme une séquence temporelle, nous pouvons analyser les coefficients de filtres dans son domaine fréquentiel (par exemple on zoome sur les coefficients du deuxième banc de filtres dans la figure 5.1). Avant d'effectuer l'analyse du domaine fréquentiel, une normalisation de la variance et de la moyenne (*Mean and Variance Normalization*, MVN) [Jain, 2001] est faite pour normaliser les moyennes des trajectoires à 0 et les écarts à 1. Comme les segments de parole ont des longueurs différentes, nous pouvons appliquer la transformation rapide de Fourier (FFT) sur les coefficients du banc de filtres, dans lesquels les longueurs des fenêtres analysées ont été fixées tel qu'indiqué dans la figure. Cette méthode est similaire à la transformation de Fourier à court terme (*short-time Fourier tranform*) utilisée pour produire les spectrogrammes. La longueur et le déplacement (*shift*) des fenêtres peuvent être déterminés empiriquement. Les spectres qui représentent les fenêtres de coefficients d'un banc de filtres (*Filterbank*) sont appelés les spectres de modulation du signal de parole correspondant à ce banc de filtres. Ces spectres capturent les informations sur les changements d'énergie dans chaque banc de filtre. Par conséquent, nous avons une séquence de spectres de modulation pour chaque banc de filtres. Comme nous voulons utiliser un vecteur de longueur fixe pour représenter un segment de parole, nous pouvons prendre la moyenne des spectres de modulation. Au final, il y a un seul spectre de modulation (le spectre moyen) pour chaque banc de filtres (*Filterbank*). La discrimination entre parole native et non native peut être réalisée en utilisant les caractéristiques extraites de ce spectre de modulation en moyenne.

D'après nos études, ce ne sont pas toutes les fréquences de modulation spectrale qui sont utiles pour la discrimination entre parole native et non native. Si nous générons 100 trames par seconde pour les bancs de filtres, nous avons 100 points de données par seconde dans les

trajectoires de bancs de filtres et donc les fréquences de modulation ont une gamme de 0 à 50Hz selon la théorie de l'échantillonnage de Shannon [Shannon, 1949]. Les chercheurs ont montré que, parmi les fréquences de modulation de 0 à 50 Hz, les fréquences 1-16 Hz sont les plus importantes pour la parole, alors que les très faibles (0-1 Hz) et hautes fréquences de modulation (> 16Hz) sont principalement dues à des bruits [Kanedera, 1999]. La raison est que les articulateurs du conduit vocal de l'homme ne peuvent pas se déplacer très rapidement, donc ne peuvent pas générer une modulation de fréquence très élevée. Il est donc raisonnable de s'attendre à ce que les fréquences de modulation utiles pour nos études soient situées dans la gamme de modulation 1-16Hz. La sélection des fréquences de modulation est illustrée sur la figure 5.2. Seule la gamme de fréquences de modulation basse est conservée, puis les vecteurs extraits sont concaténés en un seul vecteur qui est réduit par une méthode de réduction de dimension (analyse en composantes principales, PCA [Wood, 1987], dans notre cas).

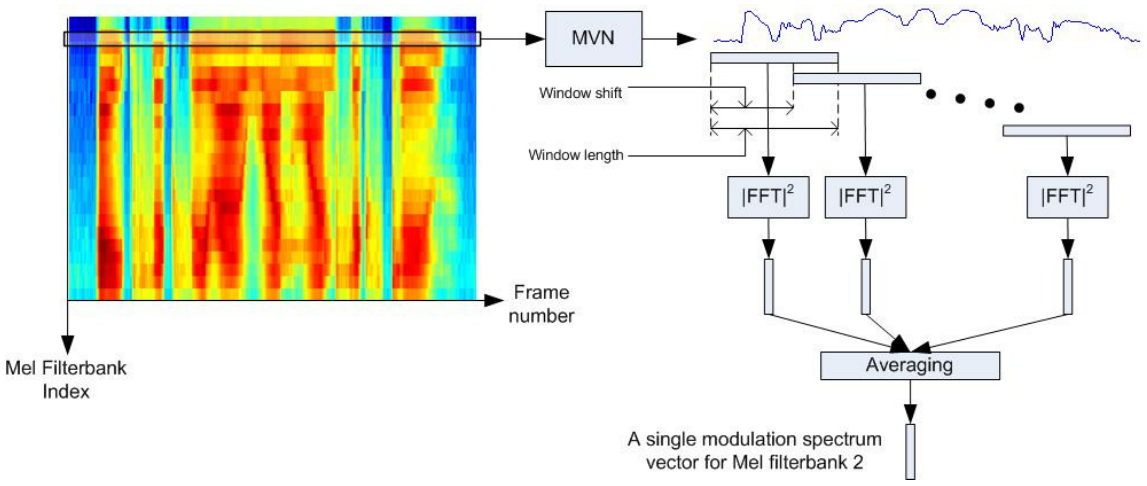



*Figure 5.1 : Exemple de génération d'une modulation spectrale pour le deuxième bloc de filtres d'un signal de parole*

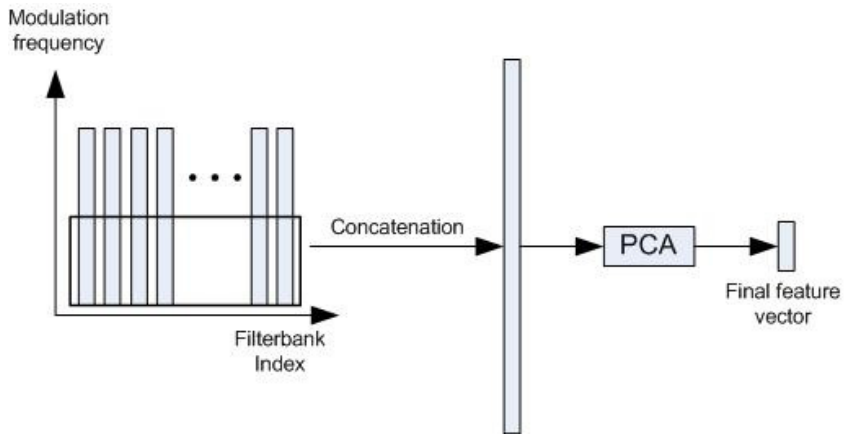



*Figure 5.2 : Extraction des caractéristiques spectrales d'un segment de parole à partir d'une séquence de modulations spectrales*

Il y a quelques paramètres à définir selon les expériences pour l'extraction de caractéristiques du spectre de modulation. Les deux premiers paramètres sont la longueur et le décalage de fenêtre (*window length and window shift*). Selon les études récentes de Xiao Xiong [Xiong, 2009] sur la modulation du spectre, la longueur de la fenêtre est fixée à 50 trames (soit 0,5 seconde) et le décalage de fenêtre est fixée à 25 trames. Le troisième paramètre est le choix des fréquences de modulation, tel qu'illustré dans la zone rectangulaire du premier panneau de la figure 5.2. Dans notre étude, nous avons constaté que les 12 premiers points parmi les 32 points de FFT (correspondant à 0-18.75Hz) contiennent des informations utiles pour notre tâche de discrimination. Le dernier paramètre, la dimension des vecteurs caractéristiques de PCA projeté, est fixé empiriquement à 10.

### 5.3.3. Pitch (fréquence fondamentale) et formants

Le pitch ou la fréquence fondamentale (F0) est une des caractéristiques prosodiques importantes du signal de la parole. L'information de pitch a été utilisée dans plusieurs systèmes de discrimination entre parole non native [Grover, 1987]. Il est montré dans [Arslan, 1996] que le contour de pitch est différent entre les locuteurs anglais, chinois, turcs et allemands (figure 5.3).

Dans cette étude, nous avons extrait la valeur de pitch du signal pour chaque segment de parole à l'aide du logiciel Praat [Boersma, 2001]. Comme il y a différentes distributions de valeurs du pitch d'un groupe de locuteurs à l'autre, par exemple les valeurs moyennes du pitch des locuteurs masculins et féminins peuvent être très différentes, nous appliquons une normalisation *MVN* (*mean and variance normalization*) [Jain, 2001], qui est couramment utilisée pour augmenter la robustesse des paramètres (minimiser les différences caractéristiques du signal entre les locuteurs et aussi les sexes) de reconnaissance vocale, sur les contours du pitch. Ensuite, la variation et l'accélération des contours du pitch qui représentent la dynamique du pitch sont calculées de la même manière que dans le cas des MFCC (section 5.3.1). Par conséquent, chaque vecteur qui représente les caractéristiques du pitch a pour dimension trois.

Une autre caractéristique importante qui est aussi utile pour l'identification des accents (selon [Grover, 1987] et [Arslan, 1996]) est constituée par les formants. Dans nos études de discrimination, les valeurs des deux premiers formants (F1 and F2) sont utilisées et sont extraites automatiquement du signal en utilisant le logiciel Praat [Boersma, 2001] (la méthode d'estimation de formant LPC [Atal, 1979] est utilisée pour cette tâche d'extraction). La variation et l'accélération qui représentent les dynamiques du formant sont aussi calculées de

la même manière que ceux du pitch. En conséquence, la dimension d'un vecteur de caractéristiques des formants est six.

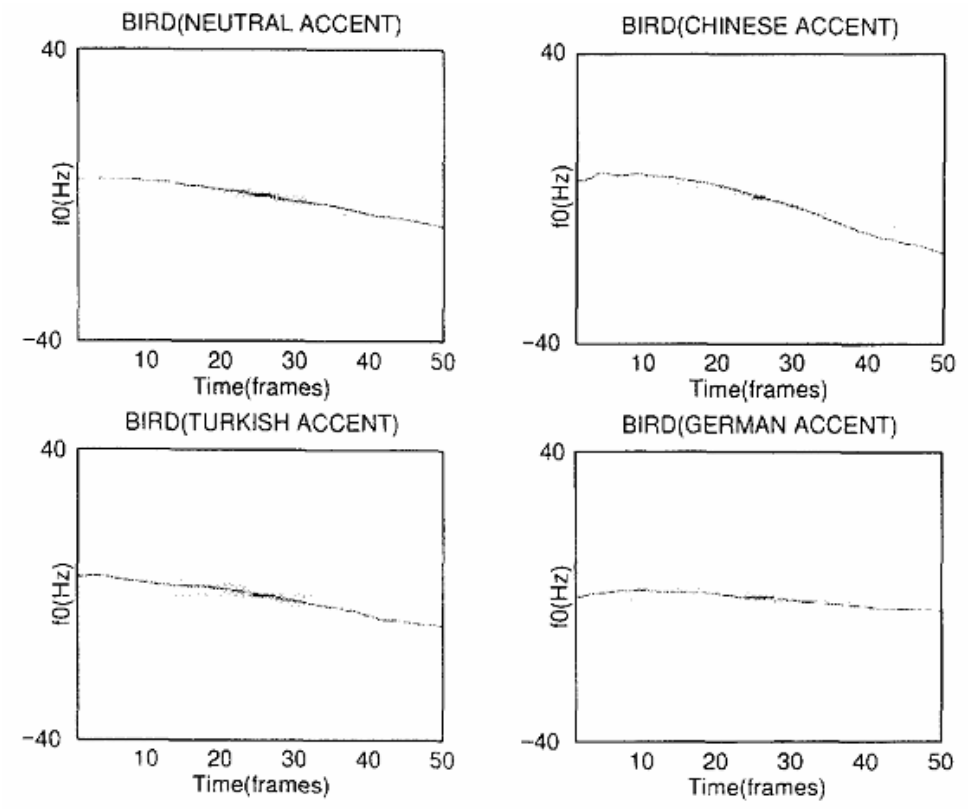



*Figure 5.3 : Histogrammes des contours du pitch normalisée (MVN) pour quatre différents accents pour le mot « BIRD » [Arslan, 1996]*

### 5.3.4. PR-VSM

Nous étudions aussi le système de discrimination entre parole native et non native à base du modèle phonotactique PR-VSM que nous avons présenté dans le chapitre 3 du manuscrit. Notre motivation à utiliser ce modèle pour la discrimination réside dans le fait que les locuteurs natifs peuvent produire des séquences phonétiques différentes de celles produites par les locuteurs non natifs.

Dans le processus de discrimination (entraîner et tester le modèle), la parole native et la parole non native sont décodées avec un système acoustico-phonétique multilingue (DAP-Mult) qui contient les unités acoustiques de trois langues (français, anglais et vietnamien). Les séquences phonétiques (les hypothèses de DAP-Mult) sont transformées en un vecteur documentaire contenant les modèles bi-grammes de phonèmes et, enfin, la technique de classification SVM (*support vector machine*) est utilisée pour classer ces vecteurs documentaires, en l'un des deux groupes de parole (native ou non native). Ce processus est le même que celui utilisé dans le module d'observateur de langue (chapitre 3) sauf que, cette

fois, nous ne générons que deux scores postérieurs pour un segment de test (un score pour la parole native et un autre pour la parole non native).

### 5.3.5. Fusion des modèles

Nous avons également fusionné les systèmes de discrimination individuels, notamment les deux meilleurs systèmes de discrimination, pour voir si le système fusionné donne des résultats de discrimination meilleurs que ceux des deux premiers systèmes. Pour chaque système, les résultats de segments de test sont récupérés, y compris les scores des segments de test natifs et non natifs. Comme les gammes des scores dans les différents systèmes peuvent être très différentes, la variance des scores est normalisée en divisant les scores par leur écart-type. Après la normalisation de la variance, pour chaque segment de test, un score fusionné est calculé comme la moyenne des scores des systèmes individuels.

## 5.4. Expérimentations

### 5.4.1. Caractéristiques des modèles de discrimination

Tous les systèmes de discrimination sont modélisés par un modèle de mélanges de gaussiennes (GMM), sauf le système PR-VSM qui utilise le modèle SVM (*Support Vector Machine*) pour classer deux classes de segments (une classe pour la parole native et une autre pour la parole non native). Pour tous les systèmes de discrimination qui utilisent des GMMs, deux GMMs sont utilisés dans chaque système, un pour la parole native et l'autre pour la parole non native. Le nombre de gaussiennes d'un GMM et le type de matrice de covariance (pleine ou diagonale) pour chaque système sont définis empiriquement, comme indiqué dans le tableau 5.2.

| **Noms des systèmes de discrimination** | **Nombre de gaussiennes d'un GMM** | **Matrice de covariance du GMM** | **Dimension** |
|---|---|---|---|
| Modulation | 4 | Pleine | 10 |
| Pitch | 64 | Diagonale | 3 |
| Formant | 64 | Diagonale | 6 |
| MFCC | 1024 | Diagonale | 36 |

*Tableau 5.2 : Configuration des modèles de mélanges gaussiens*

### 5.4.2. Métriques d'évaluation

Dans la phase de test de la performance des systèmes de discrimination, le score d'un segment de test est, en fait, un rapport de vraisemblance (*likelihood ratio, LLR*) obtenu pour les vecteurs caractéristiques de ce segment selon les GMM de parole non native et native. Ce score est calculé comme suit :

$$LLR(i)=log(p(O_i|GMM_{non-native})/p(O_i|GMM_{native})) \quad (5.1)$$

où $O_i$ est la fonction des vecteurs caractéristiques de i$^{ème}$ segment de test ; $p(O_i|GMM_{non-native})$ et $p(O_i|GMM_{native})$ sont les scores postérieurs de $O_i$ étant donné le GMM modélisant la parole non native et native, respectivement.

Dans le système PR-VSM qui utilise la technique de classification SVM au lieu de GMM, pour réaliser une classification d'un segment de test, une analyse discriminante linéaire est utilisée pour trouver une surface de décision, qui est un hyperplan entre les deux classes (la classe de la parole native et celle de la parole non native) dans la phase d'entraînement. Ainsi, le score d'un segment de test est la distance signée (négative : pour non native ; positive : pour native) du vecteur documentaire qui représente ce segment de test à la surface de décision (hyperplan). Il est calculé par :

$$f(D_i)=a^T\psi(D_i)+b \quad (5.2)$$

où $D_i$ est le vecteur documentaire du i$^{ème}$ segment de test ; $f(D_i)$ est la distance signée (négative ou positive) entre $D_i$ et la surface de décision $a^T\psi(D_i)+b=0$.

Les informations sur la formulation des GMM et SVM sont détaillées dans [Young, 1999] et [Burges, 1998; Duda, 2000] respectivement. Une comparaison entre ces deux techniques de classification statistique est présentée dans [Nazari, 2008].

Lors du processus de discrimination, notre objectif est de détecter si un segment de test est parlé par un locuteur non natif français. Cela nécessite de choisir un seuil de discrimination. Si le score du segment de test est supérieur à un seuil prédéfini, ce segment de parole est considéré comme étant prononcé par un locuteur non natif. Dans le cas contraire, le segment est considéré comme un énoncé de locuteur non natif.

Nous utilisons deux types de mesure pour mesurer la performance des systèmes de discrimination, le taux de fausses alarmes (*false alarm rate*), c'est-à-dire le pourcentage des segments de parole native qui sont détectés comme de la parole non native, et le taux de non-détection (*missing rate*), à savoir le pourcentage des segments de parole non native qui ne sont pas détectés comme des segment de parole non native. Les deux mesures dépendent du seuil

choisi. Dans notre étude, ces deux types de mesure, avec les différents seuils de discrimination de chaque système, sont tracés en utilisant les courbes DET (*detection error tradeoff, DET*) [Martin, 1997a].

Une autre mesure d'évaluation est le taux d'égale erreur (*equal error rate, EER*), à savoir le cas où le taux de non-détection est égal au taux de fausses alarmes.

### 5.4.3. Résultats d'expérimentation

Nous examinons d'abord les taux d'erreur de détection EER obtenus par les systèmes individuels et par notre essai de fusion entre les deux meilleurs systèmes. Dans le tableau 5.3, les chiffres de la deuxième colonne présentent les EER des systèmes de détection sur les BD1 et BD2. Ces chiffres représentent les taux d'erreur globaux des 15 validations croisées (*cross validation*). Par contre, les chiffres de la troisième colonne indiquent les taux d'erreur des systèmes de détection qui utilisent les BD1 et BD2 comme données d'entraînement et la BD3 comme données de test. D'après les résultats, les EER obtenus sur les données de test extraites de BD3 sont toujours plus grands que ceux obtenus sur les 15 validations croisées (BD1 et BD2) pour tous les systèmes de détection.

| **Systèmes de détection** | **EER de tâche 1 (T1)** | **EER de tâche 2 (T2)** |
|---|---|---|
| MFCC | 13,9 | 51,3 |
| Pitch | 17,8 | 22,4 |
| Formant | 19,5 | 26,2 |
| PR-VSM | 7,9 | 35,1 |
| Modulation | 7,7 | 22,7 |
| Fusion[13] | 5,1 | 13,1 |

*Tableau 5.3 : Taux d'ERR (%) des différents systèmes de détection (évalué sur les BD1, BD2 et BD3)*

En comparant ligne par ligne les EER de la tâche d'évaluation T1 (celles du BD1 et BD2) et de la tâche d'évaluation T2 (celles du BD3), nous constatons que les systèmes fondés sur les caractéristiques de pitch sont les plus robustes par rapport aux différences de canal d'enregistrement, avec la plus petite dégradation des EER (4,6 %). Les caractéristiques des formants sont également plus robustes, avec une dégradation de seulement 6,7 %. Les

[13] Nous faisons la fusion entre deux systèmes de meilleure performance de détection pour chaque tâche d'évaluation (T1 et T2).

caractéristiques de modulation sont moins robustes, avec un EER multiplié quasiment par trois. Toutefois, l'EER est seulement de 22,7 %, ce qui est le deuxième meilleur score, dans la tâche d'évaluation T2, parmi les cinq types de systèmes de discrimination. Les systèmes fondés sur MFCC et PR-VSM sont les moins robustes. Cela s'explique notamment en raison de la forte sensibilité des caractéristiques MFCC aux distorsions de canal d'enregistrement. En outre, comme la partie reconnaissance phonétique (PR) du PR-VSM utilise les caractéristiques MFCC dans son processus de décodage, PR-VSM utilise également indirectement des paramètres MFCC peu résistants aux distorsions.

Les courbes DET sur les données de test de BD3 sont tracées sur la figure. 5.4. Dans cette figure, les courbes des systèmes fondés sur le pitch et la modulation spectrale indiquent les meilleurs résultats de détection. En outre, la fusion des deux systèmes (pitch et modulation) produit des résultats optimisés, comme indiqué dans cette figure (sa courbe DET est plus proche de l'origine que les autres courbes) et aussi dans la dernière ligne du tableau 5.3 (EER = 13,1 %). En fait, selon les expérimentations, la fusion des systèmes de pitch et de modulation produit la meilleure performance parmi toutes les combinaisons possibles (nous ne présentons pas les autres combinaisons dans la figures 5.4).

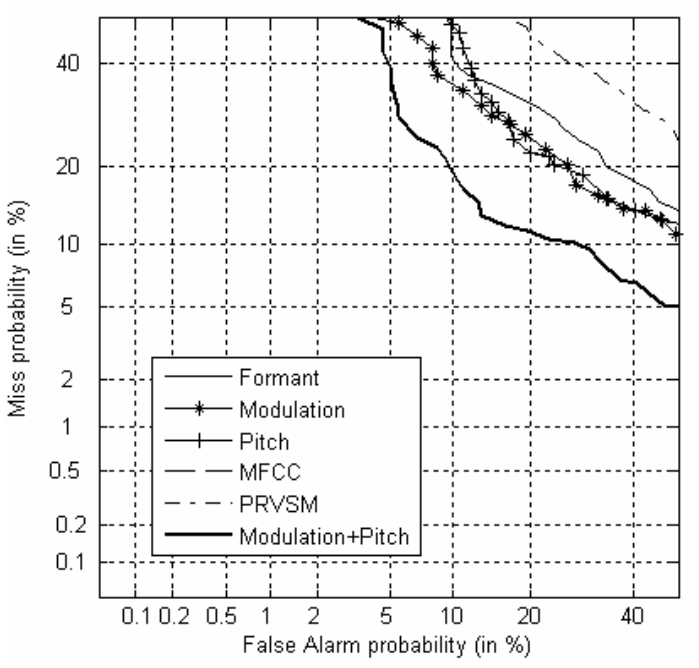



*Figure 5.4 : Courbe DET des différents systèmes de détection (en utilisant BD3 comme données de test)*

Nous allons maintenant étudier le graphique des scores normalisés sur la figure 5.5. Le graphique du haut de la figure illustre tous les scores des segments de test de parole native française extraits de la base BD3 (locuteurs français Eric, Geneviève et Mathias). Les noms

des locuteurs natifs de la langue française sont imprimés juste au-dessus de l'axe horizontal (axe des abscisses). Chaque point de la courbe représente un segment de test. Les variances de chaque score sont toutes normalisées à la valeur unité pour une meilleure comparaison. En outre, une ligne horizontale en pointillés est tracée pour chaque système afin de représenter l'axe de décision (non native ou native) fondé sur un seuil. Le seuil de chaque système de détection qui présente l'axe de décision est le point d'erreur égale (le point où le taux de non-détection est égal au taux de fausses alarmes). Ces seuils sont différents d'un système à l'autre (par exemple, le seuil du système de détection PR-VSM est 0,3 tandis que celui du MFCC est -0,8). Le graphique inférieur de la figure montre les scores des locuteurs non natifs parlant français (Andrew (Anglais), Diep, Hien, Khoa et Son (Vietnamiens), et Sethserey (Cambodgien)). Les caractéristiques de ce dernier graphique sont les mêmes que celles du graphique supérieur de la figure.

Sur la figure 5.5, d'un point de vue global, nous constatons qu'aucun système de discrimination fondé sur MFCC et PR-VSM ne permet de différencier très clairement parole native et non native. En effet, ces deux détecteurs sont peu robustes vis-à-vis des distorsions entre les données d'entraînement (DB1 et DB2) et les données de test (DB3). Au contraire des systèmes de détection MFCC et PR-VSM, les systèmes de discrimination fondés sur la modulation de spectre, sur les formants et sur le pitch sont plus efficaces, sauf pour le locuteur « Geneviève » (locuteur féminin français). Une explication possible est qu'il y a très peu de locuteurs natifs féminins dans le corpus d'entraînement des systèmes de détection. Si nous observons les bases de données d'entraînement (BD1 et DB2), parmi les 16 locuteurs non natifs, il y a huit locuteurs féminins et huit locuteurs masculins. Au contraire, il n'y a que quatre locuteurs féminins français parmi les 15 locuteurs natifs français. Cependant, le système à fusion (Modulation + Pitch) classe assez bien la parole native de ce locuteur (Geneviève). En revanche, de nombreuses erreurs de classification se produisent pour les locuteurs non natifs. Par exemple, nous observons que les scores du locuteur « Andrew » (de langue maternelle anglaise) et du locuteur « Sethserey » (de langue maternelle khmère) sont presque tous mal classés par tous les systèmes. Cela peut être dû au fait que leurs langues maternelles (l'anglais et le khmer) ne sont pas présentes dans les données d'entraînement (nous avons seulement les locuteurs vietnamiens et chinois dans nos données d'entraînement). Les autres locuteurs non natifs sont tous vietnamiens et la performance de discrimination est bonne sur ce sous-groupe.

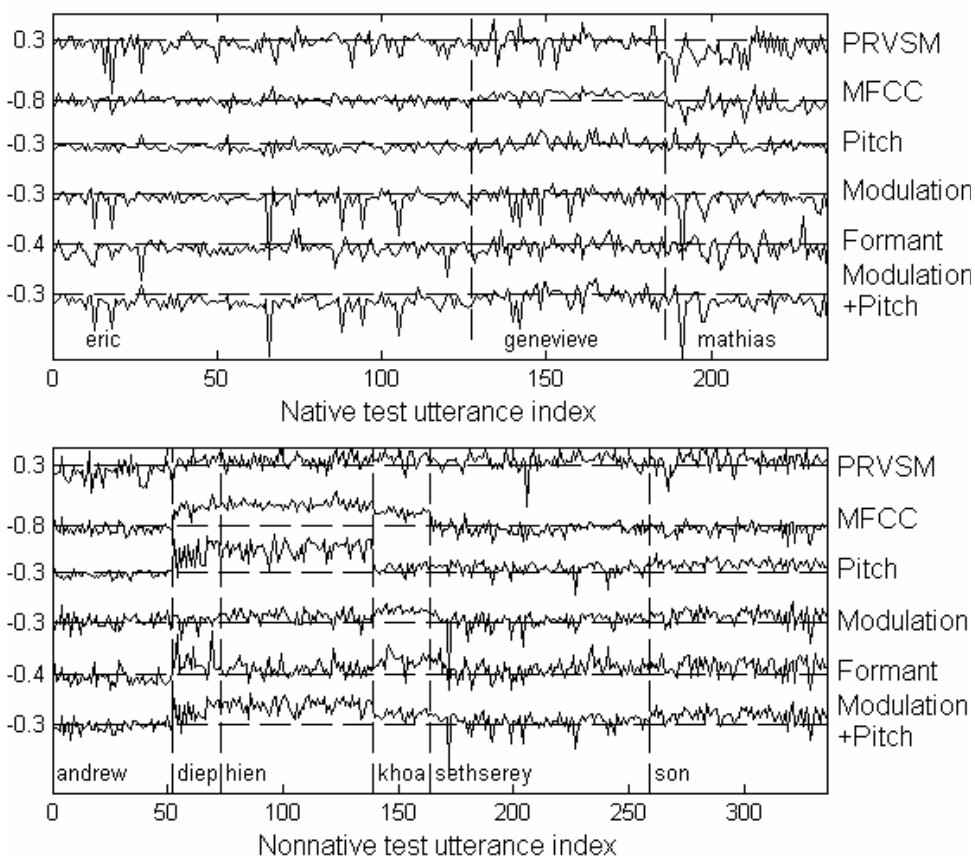



*Figure 5.5 : Graphiques des scores des données de test (BD3) pour les différents systèmes de détection*

## 5.5. Utilisation des systèmes de discrimination entre parole native et non native dans le processus d'adaptation autonome des modèles acoustiques multilingues

### 5.5.1. Discrimination entre parole native/non native suivie d'une adaptation autonome

Comme illustré figure 5.6, nous avons tenté de mettre en cascade notre meilleur système de discrimination entre parole native et non native avec notre processus d'adaptation autonome des modèles acoustiques multilingues, qui n'est alors appliqué que sur la parole non native détectée. Dans le processus d'adaptation autonome, nous utilisons l'observateur PR-VSM mentionné dans le chapitre 3 pour générer les scores postérieurs des langues, et l'approche INTER-MLLR (chapitre 4) est utilisée pour adapter le modèle acoustique multilingue afin de mieux traiter les segments de parole non native. Au contraire, le groupe de segments de parole native française détecté par le système est décodé par le décodeur acoustico-phonétique français (DAP-Mono FR) qui utilise le modèle acoustique français

appris sur le corpus BREF [Lamel, 1991] (voir chapitre 4 pour des informations plus détaillées sur le modèle acoustique français).

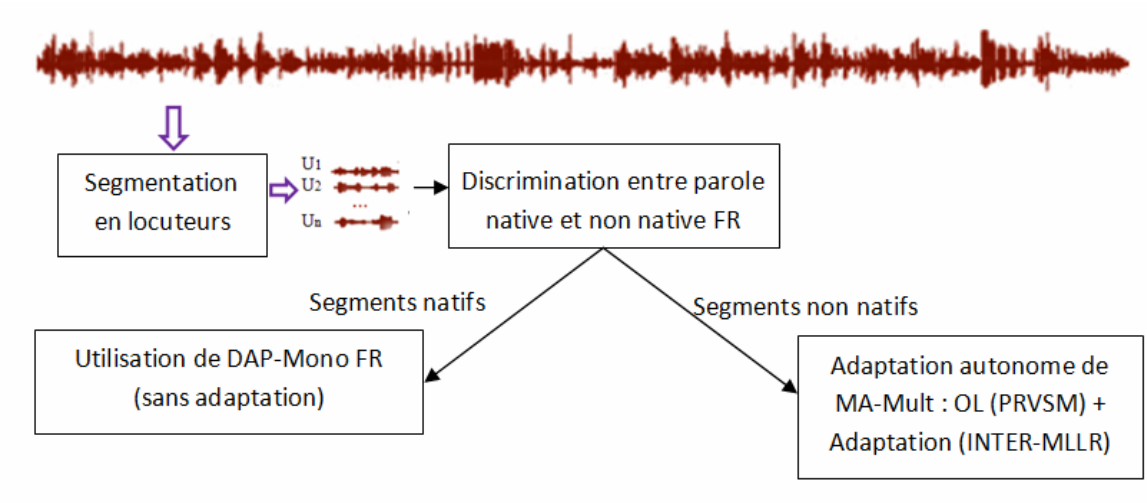



*Figure 5.6 : Utilisation d'un système de discrimination entre parole native et non native française en amont du processus d'adaptation autonome du modèle acoustique multilingue*

### 5.5.2. Résultats d'expérimentation

Comme les systèmes de discrimination sont entraînés sur de la parole native et non native du français, nous n'extrayons que la parole native et non native française, comme mentionnée dans les données de test (tableau 2.3 du chapitre 2), pour étudier si le nouveau processus d'adaptation proposé (détection + adaptation autonome) améliore la performance du décodage phonétique de la parole native et de la parole non native.

| **Origine** | Français | Vietnamien | Anglais | Total |
|---|---|---|---|---|
| **Quantité** | *241*<br>(3) | *486*<br>(3) | *253*<br>(1) | *980*<br>(7) |

*Tableau 5.4 : Quantité de parole en français (chiffres en italique sont les valeurs en seconde) et nombre de locuteurs (chiffres entre parenthèses) extraits du corpus de test pour l'adaptation autonome*

#### *5.5.2.1. Résultats de la discrimination (sur le sous-corpus en français)*

Selon les résultats de détection mentionnés dans la section 5.4, les systèmes de détection fondés sur les coefficients cepstraux MFCC et l'approche phonotactique PR-VSM sont les moins performants parmi les cinq systèmes de détection étudiés. Pour cette raison, ces deux systèmes de détection ne sont pas présentés dans nos études d'adaptation des modèles acoustiques. Nous analysons les résultats de détection des différents systèmes en fonction du taux d'erreur égale (*EER*), présenté dans le tableau 5.5.

| **Système** | **Formant** | **Modulation** | **Pitch** | **Fusion (Modulation+Pitch)** |
|---|---|---|---|---|
| EER ( %) | 38,6 | 25,5 | 28 | 5,6 |

*Tableau 5.5 : Taux d'erreur de détection des différents systèmes pour la parole non native française*

Selon les taux d'erreur de détection (tableau 5.5), les systèmes de discrimination fondés sur le pitch et la modulation sont les deux meilleurs systèmes de discrimination entre parole native et parole non native française. Ces résultats sont cohérents avec ceux des systèmes de discrimination qui utilisent le corpus complet comme données de test (section 5.4.3). La fusion entre les deux meilleurs systèmes (pitch et modulation) donne la meilleure performance de détection. Curieusement, la fusion entre les trois meilleurs systèmes de détection (pitch, modulation et formant) dégrade légèrement la performance du système (le taux d'erreur de cette dernière fusion n'est pas présenté dans le tableau 5.5).

#### *5.5.2.2. Résultats d'adaptation des modèles acoustiques multilingues*

Dans cette section, nous observons les résultats de différents systèmes de décodage acoustico-phonétique.

- DAP-Mult (*baseline*) : tous les segments de parole (native et non native) française sont décodés avec le décodeur multilingue de référence qui utilise le modèle acoustique multilingue à trois langues (EN, FR et VN) mentionné dans la section 3.3.2 du chapitre 3.
- Autonome : tous les segments (natifs et non natifs) sont décodés par le système de décodage phonétique qui utilise le modèle acoustique multilingue adapté. Ce modèle adapté est créé en utilisant l'adaptation autonome (l'observateur de langue PR-VSM suivi de l'adaptation INTER-MLLR) qui est déjà détaillé dans les chapitres précédents.
- Détection + autonome : dans ce cas, le module de discrimination entre parole native et non native classe tous les segments à décoder en deux groupes (natif et non natif) ; les segments natifs sont décodés par le décodeur acoustico-phonétique français (DAP-Mono FR) tandis que les segments non natifs sont décodés par le système de décodage phonétique fondé sur l'adaptation autonome (Autonome). Nous étudions aussi le cas parfait de discrimination entre parole native et non native (Oracle + Autonome).

Le tableau 5.6 présente les taux d'erreur des décodeurs acoustico-phonétiques testés. En observant les résultats présentés dans le tableau 5.6, nous trouvons que les performances des décodeurs qui utilisent les systèmes de discrimination entre parole native et non native avant le processus de décodage ou d'adaptation sont généralement meilleures que celles du décodeur multilingue de référence et celles du décodeur avec adaptation autonome. L'amélioration dépend de la qualité du système de détection (par exemple, mettre en cascade le système de détection fusionné avec notre processus d'adaptation autonome donne les meilleurs résultats).

| | **DAP-Mult (*baseline*)** | **Autonome** | **Formant+ Autonome** | **Pitch+ Autonome** | **Modulation+ Autonome** | **Fusion+ Autonome** | **Oracle+ Autonome** |
|---|---|---|---|---|---|---|---|
| **FRfr** | 56,7 | 57,8 | 52,4 | 52,6 | 52,5 | 52,4 | 52,2 |
| **FRen** | 59,1 | 51,3 | 54,1 | 53,1 | 52,8 | 51,4 | 51,3 |
| **FRvn** | 55,8 | 55,0 | 54,9 | 54,5 | 54,9 | 55,0 | 55,0 |
| **Moyenne** | 57,2 | 54,7 | 53,9 | 53,7 | 53,5 | 53,1 | 52,9 |

*Tableau 5.6 : Taux d'erreur de phonèmes (%) de décodage phonétique des différents systèmes acoustico-phonétiques (adapté et non adapté)*

## 5.6. Conclusion

Nous avons présenté dans ce chapitre nos études sur plusieurs systèmes de discrimination entre parole native et non native en français.

Selon les résultats de l'expérimentation, les systèmes fondés sur la modulation spectrale et les caractéristiques phonétiques comme le pitch et les formants sont les plus robustes pour notre tâche de discrimination surtout quand il y a des distorsions de canal entre les données d'entraînement et les données de test. Par contre, les systèmes de discrimination fondés sur les coefficients cepstraux MFCC et ceux fondés sur l'approche phonotactique PR-VSM ne sont pas recommandés s'il existe des disparités de canal entre les données d'entraînement et les données de test.

Pour finir, nous avons vérifié que mettre un tel système de discrimination entre parole native et non native en amont du processus d'adaptation permet d'améliorer les performances du décodage acoustico-phonétique, la parole non native étant traitée selon notre technique d'adaptation autonome, et la parole native étant transcrite par un décodeur monolingue.

# Conclusion et perspectives

## Conclusion

Le travail présenté dans ce mémoire porte sur l'amélioration des modèles acoustiques multilingues pour la reconnaissance automatique de la parole de type « réunions multilingues ». L'étude de ce type de parole pose plusieurs problèmes difficiles : 1) dans les réunions multilingues, il peut y avoir de la conversation entre locuteurs natifs et non natifs ; 2) il y a de la parole non native en plusieurs langues parlées par des locuteurs venant de différentes origines ; 3) il n'y a pas suffisamment de données pour amorcer les systèmes de reconnaissance pour ce type de parole, etc. Pour répondre à ces défis, nous proposons une adaptation des modèles acoustiques multilingues qui s'appelle « l'adaptation autonome » dans laquelle les données d'adaptation ne sont pas disponibles et les langues parlées et maternelles des locuteurs sont supposées inconnues au départ (adaptation non supervisée). D'un point de vue plus opérationnel, nous utilisons une adaptation autonome du modèle acoustique multilingue pour améliorer la performance d'un système de transcription phonétique de parole de type « réunion multilingue » dans lequel la parole native et non native dans trois langues est étudiée : anglais (EN), français (FR) et vietnamien (VN).

La première partie de cette thèse a été consacrée à construire le corpus pour notre étude. En ce qui concerne le recueil de parole native et non native à partir des réunions multilingues, nous avons présenté notre corpus d'enregistrement de réunions virtuelles appelé MICA-MultiMeet. À la fin du processus d'enregistrement, le corpus contient de la parole native et non native pour trois langues (français, anglais et vietnamien) et de la parole native pour la langue khmère (nous avons eu des difficultés à trouver les locuteurs étrangers pour la langue khmère). Ce corpus représente environ six heures de signal de parole (cinq heures transcrites et une heure non transcrite). La parole non native représente la majorité du corpus (66 % du signal transcrit sont de la parole non native). Des données de test ont été extraites à partir du corpus précédent pour nos expérimentations concernant l'adaptation autonome des modèles acoustiques multilingues. À cause de l'absent de parole non native en khmer, le corpus de test ne contient que de la parole native et non native en français, anglais et vietnamien.

La deuxième partie de la thèse a été consacrée aux méthodes d'adaptation autonome de modèles acoustiques multilingues. Les méthodes d'adaptation proposées doivent surmonter deux défis importants : 1) les langues parlées et les origines (langues maternelles) des locuteurs sont inconnues au moment du traitement ; 2) il n'y a aucune donnée disponible pour l'adaptation. Pour relever ces défis, le processus d'adaptation autonome proposé contient deux modules principaux.

Le premier module s'appelle « l'observateur de langues » et consiste à récupérer les caractéristiques linguistiques (les langues parlées et les origines des locuteurs) des segments à décoder. Nous avons présenté différentes approches pour construire un observateur de langues qui génère des scores postérieurs (pour chaque langue) des langues impliquées (anglais, français et vietnamien). Nous avons vérifié l'hypothèse selon laquelle un tel observateur peut fournir des informations utiles pour l'adaptation non supervisée d'un modèle acoustique multilingue lors du décodage d'un segment de parole pour lequel la langue et l'origine du locuteur sont inconnues. Selon les résultats d'expérimentation, l'observateur de langue fondé sur l'approche phonotactique PR-VSM, qui a été récemment proposée dans les études de reconnaissance des langues, semble le plus efficace, et est considéré par la suite dans notre processus d'adaptation non supervisée.

Le deuxième module consiste à adapter le modèle acoustique multilingue. Les méthodes proposées dans ce dernier module ne demandent que les connaissances fournies par l'observateur de langue dans son processus d'adaptation du modèle acoustique multilingue, sauf dans le cas de l'adaptation MLLR. Toutes les techniques d'adaptation n'utilisent aucune donnée supplémentaire dans le processus d'adaptation. Nous avons étudié deux types d'adaptation non supervisée d'un modèle acoustique multilingue. Le premier type considère, dans son processus d'adaptation, le segment de parole en cours de décodage ainsi que les segments déjà décodés (les segments dans l'historique du décodage). Ce type d'adaptation est dit « hors ligne » (*offline multilingual acoustic model adaptation*). Contrairement à l'adaptation « hors ligne », le deuxième type, qui est l'adaptation « en ligne » (*online multilingual acoustic model adaptation*), ne considère que le segment de la parole en train d'être décodé. Selon les résultats d'expérimentation, les adaptations « hors ligne » présentent des résultats légèrement meilleurs que les adaptations « en ligne » pour les segments de parole native. Au contraire, les adaptations « en ligne » sont recommandées pour les segments de parole non native. Les techniques d'adaptation « en ligne » que nous avons proposées, appelées « interpolation des modèles acoustiques » (INTER et INTER-MLLR) fondées sur

l'observateur de langues PR-VSM, sont des choix prometteurs pour traiter la parole non native. En revanche, elles dégradent les performances sur de la parole native.

Pour rendre le processus de décodage phonétique plus robuste à la fois pour la parole native et non native, nous avons proposé d'utiliser un module de discrimination entre parole native et parole non native en début du processus d'adaptation autonome. Nous avons étudié plusieurs approches à cette discrimination, proposées dans des études récentes sur la détection des accents de locuteurs étrangers, comme les approches fondées sur le pitch, qui est une des caractéristiques les plus importantes de la prosodie du signal, sur les deux premiers formants du signal (F1 et F2), sur les caractéristiques cepstrales (MFCC et la modulation du spectre), et aussi sur le modèle phonotactique fondé sur les séquences phonétiques du décodage (PR-VSM). Dans notre étude, nous avons appliqué les systèmes de détection fondés sur ces approches à la discrimination entre parole native et non native de la langue française. Selon les résultats d'expérimentation, les systèmes fondés sur la modulation spectrale, le pitch et les formants sont les plus robustes pour notre tâche de discrimination, surtout quand il y a des distorsions entre les données d'entraînement et les données de test. Au contraire, les systèmes de discrimination fondés sur des coefficients cepstraux MFCC et sur l'approche phonotactique PR-VSM ne sont pas recommandés s'il y a une disparité entre les données d'entraînement et les données de test. En outre, quand nous couplons ces systèmes de discrimination avec le processus d'adaptation autonome des modèles acoustiques multilingues, nous trouvons que les performances des décodeurs sont meilleures que celles des décodeurs sans système de discrimination.

En conséquence, les avantages de l'adaptation autonome de modèles acoustiques multilingues sont les suivants :

- elle est efficace pour traiter de la parole non native pour laquelle les langues secondes et les langues maternelles des locuteurs sont inconnues (adaptation non supervisée) ;
- elle n'utilise aucune donnée supplémentaire dans son processus d'adaptation, ce qui est nécessaire car, en général, la collecte de données pour adapter un modèle acoustique à toutes les langues et toutes les origines n'est pas envisageable.

Comme l'adaptation autonome fondée sur les techniques d'interpolation des modèles acoustiques (INTER et INTER-MLLR) n'est pas recommandée pour la parole native, car elle augmente légèrement les taux d'erreur, l'utilisation du module de discrimination entre parole native et non native au début de la chaîne d'adaptation autonome permet d'optimiser la

performance globale du système sur un corpus contenant de la parole native et de la parole non native.

# Perspectives

Plusieurs travaux sont envisagés dans la continuité de notre travail.

Premièrement, la collecte de plus de données de parole native et non native est prévue. Cette tâche devait être réalisée dans le cadre du projet PI[14] par l'un des partenaires du projet, le centre de recherche MICA à Hanoi, en collaboration avec l'Institut de Technologie du Cambodge (ITC) à Phnom Penh et le laboratoire LIG à Grenoble. L'objectif est de construire un grand corpus de même nature que le corpus « MICA-MultiMeet » pour plusieurs langues (notamment les langues peu dotées) et plusieurs origines de locuteurs. Ce corpus pourra être utilisé non seulement pour les travaux de recherche en traitement des langues naturelles, comme la reconnaissance ou la synthèse de la parole, mais aussi pour des travaux linguistiques comme l'analyse des accents des locuteurs, l'analyse des caractéristiques prosodiques et acoustiques de la parole non native, etc.

Dans la suite de notre travail, nous souhaitons utiliser ce corpus pour trois tâches principales que nous pouvons résumer ainsi :

- appliquer les techniques d'adaptation étudiées aux langues peu dotées disponibles dans le corpus et en particulier au khmer (le corpus construit devrait contenir de la parole non native du khmer) ;
- créer des systèmes de discrimination entre parole native et non native pour d'autres langues (anglais, vietnamien, khmer, etc.).
- analyser les caractéristiques prosodiques, acoustiques, et phonétiques de la parole native et non native et les utiliser en les intégrant dans des méthodes automatiques (adaptation de modèles acoustiques multilingues ou discrimination entre parole native et non native) pour améliorer la performance des systèmes de reconnaissance.

De plus, l'amélioration du temps d'exécution de l'adaptation autonome des modèles acoustiques multilingues, surtout l'adaptation fondée sur l'interpolation de modèles acoustiques, est aussi à considérer. Nous envisageons de comparer les techniques d'interpolation « en ligne » avec les techniques d'interpolation « hors ligne » (INTER ou

[14] Le projet PI (ANR BLANC 2009-2010, http://pi.imag.fr) concerne le traitement automatique du langage parlé (notamment la reconnaissance automatique de la parole) pour les langues peu dotées. Les partenaires sont le LIG, le LIA, et le centre international MICA (Hanoï, Vietnam).

INTER-MLLR). Dans ce mémoire (chapitre 4), nous avons présenté l'interpolation des modèles acoustiques « en ligne ». Dans ce cas, la langue qui présente le meilleur score postérieur fourni par l'observateur de langue (OL) est considérée comme la langue cible et les deux autres langues sont considérées comme les langues sources dans le processus d'interpolation (les scores postérieurs fournis par l'OL étant considérés comme les poids d'interpolation). Ce processus d'interpolation est effectué pour chaque segment de parole à décoder. Au contraire, le processus d'interpolation « hors ligne » que nous pourrions étudier consisterait à classer tous les énoncés oraux à traiter, selon notre observateur de langue, en fonction de la langue parlée et des hypothèses d'origine, avant de réaliser l'adaptation.

Jusqu'à présent, nous avons utilisé, pour la discrimination entre la parole native et non native, des modèles de classification de type GMM (*Gaussian Mixture Model*) et SVM (Séparateur à Vaste Marge). Ces deux modèles ont besoin de beaucoup de données pour entraîner les systèmes de discrimination entre parole native et non native. C'est pour cette raison que nous ne pouvons pas étudier les systèmes de détection de parole native et non native pour toutes les langues impliquées dans le corpus MICA-MultiMeet. Il faut donc chercher des méthodes nécessitant moins de données d'apprentissage et pouvant être relativement (voir totalement) indépendantes de la langue parlée. Pour aller dans cette direction, il est nécessaire d'étudier plus en profondeur ce qu'est la parole non native, en analysant ses caractéristiques acoustiques, prosodiques, etc.

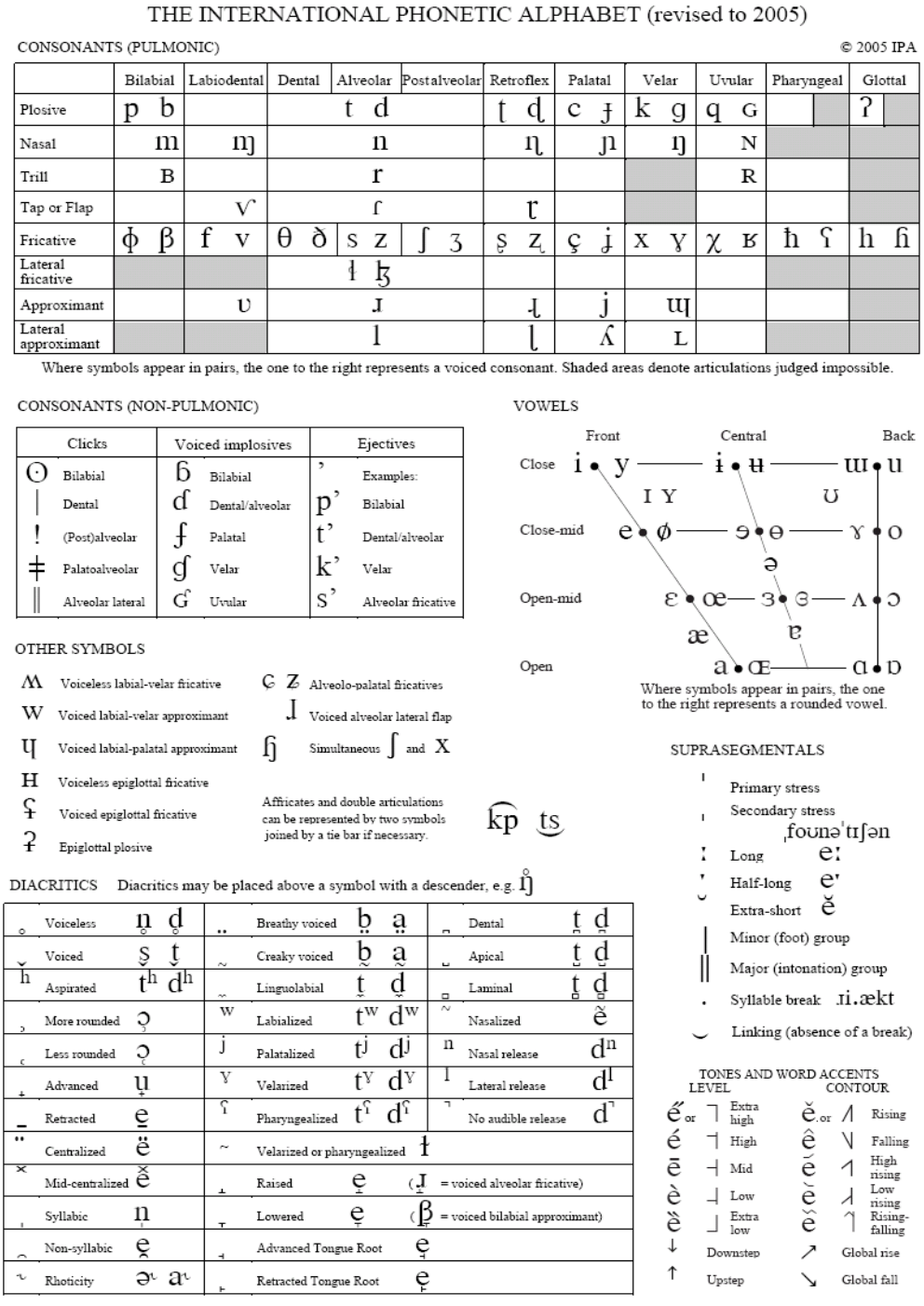
La modélisation multilingue du langage et du dictionnaire de prononciation sont aussi dans notre agenda. Les systèmes actuels génèrent, à partir de segments de parole, des séquences de phonèmes au lieu de séquences de mots. Afin de construire des systèmes adaptés qui fournissent des séquences de mots, la modélisation multilingue du langage et du dictionnaire de prononciation doivent être construits. Le concept d'autonomie (pour mieux traiter la parole non native), évalué ici dans le cadre des modèles acoustiques, pourrait ainsi être étendu aux modèles de prononciation.

Modéliser la variation dialectale dans les modèles de langue est aussi un défi ambitieux qui fait récemment l'objet de recherches dans la communauté du traitement automatique du langage naturel (voir par exemple le premier atelier « Algorithms and Resources for Modelling of Dialects and Language Varieties » lié à la conférence EMNLP[15] 2011). Ce domaine nouveau fera sans doute partie de nos recherches futures.

[15] http://www.ofai.at/~dialects2011/

# Annexe A :

# Tableau de l'API[16]




[16] http://en.wikipedia.org/wiki/International_Phonetic_Alphabet

# Annexe B :

# Tableau de l'X-SAMPA[17]

| Consonants (pulmonic) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Place of articulation → | Labial | | Coronal | | | | Dorsal | | | Radical | | | (none) |
| Manner of articulation ↓ | Bilabial | Labio-dental | Dental | Alveolar | Post-alveolar | Retro-flex | Palatal | Velar | Uvular | Pharyn-geal | Epi-glottal | | Glottal |
| Nasal | m | F | | n | | n` | J | N | N\ | | | | |
| Plosive | p b | p_d b_d | | t d | | t` d` | c J\ | k g | q G\ | | >\ | | ? |
| Fricative | p\ B | f v | T D | s z | S Z | s` z` | C j\ | x G | X R | X\ ?\ | H\ <\ | | h h\ |
| Approximant | B_o | v\ | | r\ | | r\` | j | M\ | | | | | |
| Trill | B\ | | | r | | * | | | R\ | | * | | |
| Tap or Flap | *† | *† | | 4 | | r` | | | | | * | | |
| Lateral Fricative | | | | K K\ | | * | * | * | | | | | |
| Lateral Approximant | | | | l | | l` | L | L\ | | | | | |
| Lateral Flap | | | | l\ | | * | * | * | | | | | |

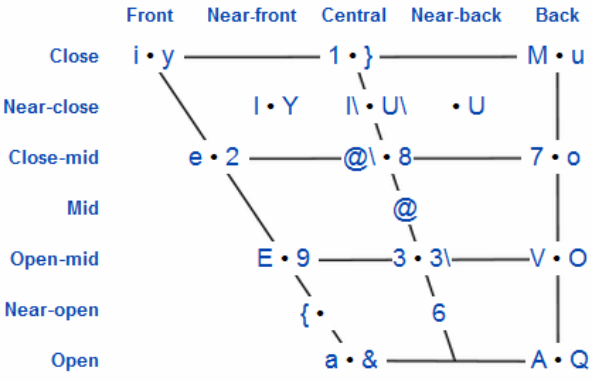
- Daggers (†) mark IPA symbols that have recently been added to Unicode. Since April 2008, this is the case of the labiodental flap, symbolized by a right-hook *v* in the IPA: ⱱ A dedicated symbol for the labiodental flap does not yet exist in X-SAMPA.

| Coarticulated | |
|---|---|
| W | Voiceless labialized velar approximant |
| w | Voiced labialized velar approximant |
| H | Voiced labialized palatal approximant |
| s\ | Voiceless palatalized postalveolar (alveolo-palatal) fricative |
| z\ | Voiced palatalized postalveolar (alveolo-palatal) fricative |
| x\ | Voiceless "palatal-velar" fricative |

| Affricates and double articulation | |
|---|---|
| ts | voiceless alveolar affricate |
| dz | voiced alveolar affricate |
| tS | voiceless postalveolar affricate |
| dZ | voiced postalveolar affricate |
| ts\ | voiceless alveolo-palatal affricate |
| dz\ | voiced alveolo-palatal affricate |
| tK | voiceless alveolar lateral affricate |
| kp | voiceless labial-velar plosive |
| gb | voiced labial-velar plosive |
| Nm | labial-velar nasal (stop) |

| Consonants (non-pulmonic) | | | | | |
|---|---|---|---|---|---|
| Clicks | | Implosives | | Ejectives | |
| O\ | Bilabial | b_< | Bilabial | _> | *For example:* |
| \|\ | Laminal alveolar ("dental") | d_< | Alveolar | p_> | Bilabial |
| ! | Apical (post-) alveolar ("retroflex") | J_< | Palatal | t_> | Alveolar |
| =\ | Laminal postalveolar ("palatal") | g_< | Velar | k_> | Velar |
| \|\|\ | Lateral coronal ("lateral") | G_< | Uvular | s_> | Alveolar fricative |

## Vowels




[17] http://en.wikipedia. org/wiki/Extended_Speech_Assessment_Methods_Phonetic_Alphabet

# Annexe C :
# Transcription automatique d'une langue en danger, le Mo Piu

## C.1. Introduction

Parmi les quelque 6000 langues[18] du monde, un grand nombre est en train de disparaître et ce phénomène s'accélère d'année en année. Selon l'organisme UNESCO (*United Nations Educational, Scientific and Cultural Organization*), ces langues sont définies comme les langues en danger appartenant au patrimoine culturel immatériel de l'humanité, et nécessitent des programmes de sauvegarde. L'UNESCO fournit les chiffres suivants :

- 50 % des langues sont en danger de disparition ;
- une langue disparaît en moyenne toutes les deux semaines ;
- si rien n'est fait, 90 % des langues vont probablement disparaître au cours de ce siècle.

L'UNESCO a notamment publié un atlas mondial des langues en danger, détaillé dans [UNESCO, 2010].

Au niveau des études sur le traitement des langues en danger, plusieurs conférences (par exemple les conférences SEALS[19], FEL[20], SLTU[21], Interspeech'10[22]) ont été mises en place afin de publier les études et recherches sur les analyses linguistiques, et de diffuser des outils de développement pour aider à la tâche de sauvegarde des langues en danger.

[18] Selon le site « Langues en danger » de l'UNESCO (nov. 2004), le nombre de 6000 langues est controversé, car il dépend de la définition de ce qu'est une langue, sur laquelle il n'y a pas consensus. La base Ethnologue donne le nombre de 6700 (nov. 2004), mais l'estimation va de 3000 à 7000 selon les sources.

[19] SEAL : Southeast Asian Linguistics Society (http://www.jseals.org/)

[20] FEL: *Foundation for Endangered Languages* (http://www.ogmios.org/conferences/)

[21] SLTU : *Spoken Languages Techologies for Under-resourced languages (*http://www.mica.edu.vn/sltu)

[22] Le terme utilisé par la conférence Interspeech'10 (Interspeech est une des plus grandes conférences en traitement de la parole) est « *spoken language processing for all* » (http://www.interspeech2010.org).

Dans le cadre du projet PI[23], nous menons des études linguistiques et développons des outils informatiques pour sauvegarder la langue d'une minorité ethnique, le mo piu, qui se rattacherait aux langues Hmong. Cette minorité est située dans les montagnes au Nord du Vietnam (près de la frontière du Vietnam et de la Chine). Selon les premières études effectuées en 2010 [Caelen-Haumont, 2010], le mo piu est vraiment une langue en danger, car elle est non répertoriée, non documentée, sans écriture, et parlée par seulement 227 personnes. De plus, aucune ethnie environnante ne comprend cette langue.

En dehors des études phonétiques et tonologiques menées par ailleurs à MICA [Caelen-Haumont, 2010] en collaboration avec d'autres laboratoires (LACITO[24]) et universités (Provence Aix-Marseille I, et Paris 7), une des méthodes pour documenter le mo piu consiste à utiliser des outils informatiques pour transcrire les segments de parole en mo piu en séquences de phonèmes ou de mots [Caelen-Haumont G., 2011]. Pour cette tâche, comme il n'y a pour cette langue ni modèle acoustique, ni modèle de langue, ni modèle lexical, nous proposons d'utiliser des modèles acoustiques d'autres langues. Dans des travaux antérieurs [Schultz, 2006], les systèmes acoustico-phonétiques multilingues sont plus souvent préférés que les systèmes monolingues pour transcrire de la parole d'une nouvelle langue qui n'a aucune donnée d'adaptation ou d'apprentissage de modèle acoustique. Les tâches que nous détaillerons dans les sections suivantes consistent à trouver quels modèles acoustiques multilingues sont les plus robustes pour la tâche de transcription du mot piu et pourront faciliter les travaux des phonéticiens dans leurs analyses acoustiques et phonétiques de la parole de cette langue minoritaire particulière.

## C.2. Processus de transcription automatique

Le processus de transcription automatique des segments de parole mo piu en séquences de phonèmes est détaillé comme suit : dans la figure C.1, les extraits en mo piu sont décodés par un système de reconnaissance acoustico-phonétique multilingue (le terme « acoustico-phonétique » signifie que le système ne dépend que des modèles acoustiques). Pour rendre les résultats plus lisibles par les phonéticiens, les hypothèses de sortie sont converties en deux fichiers TextGrid reconnus par l'outil « Praat » [Boersma, 2001]. Le premier contient les

[23] Le projet PI (ANR BLANC 2009-2010, http://pi.imag.fr) concerne le traitement automatique du langage parlé (notamment la reconnaissance automatique de la parole) pour les langues peu dotées et pas dotées. Les partenaires sont le LIG, le LIA, et le centre international MICA (Hanoï, Vietnam).

[24] http://lacito.vjf.cnrs.fr/index.htm

séquences phonétiques en format API [IPA, 1999] et le second contient les séquences phonétiques en X-SAMPA [Wells, 1997].

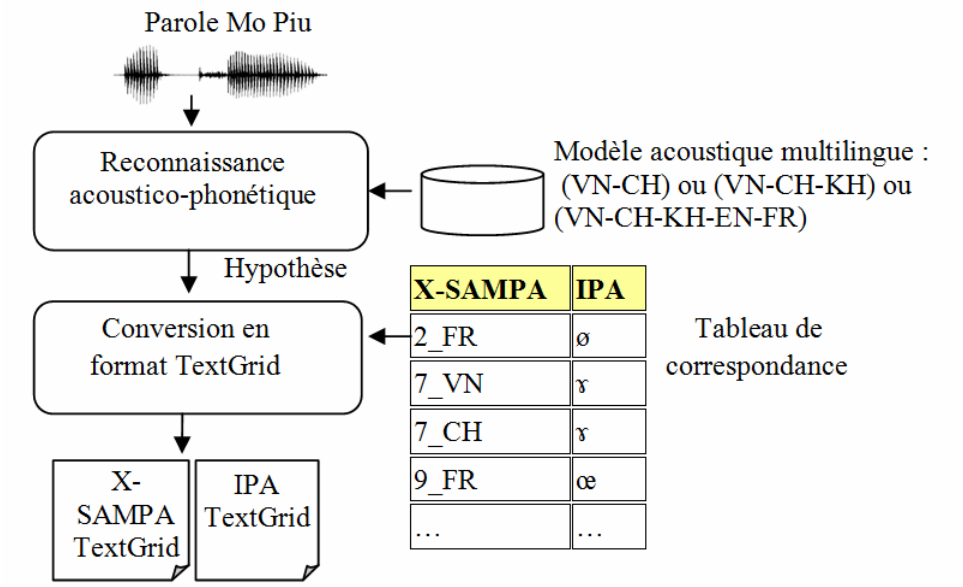



*Figure C.1 : Processus d'annotation automatique du mo piu*

Dans les sections suivantes, nous détaillons les caractéristiques des modèles acoustiques multilingues utilisés et le tableau de correspondance phonétique entre les phonèmes en format API et ceux en format X-SAMPA.

### C.2.1. Modèles acoustiques multilingues

En utilisant les modèles acoustiques monolingues disponibles dans nos travaux de thèse, nous avons créé des modèles acoustiques multilingues (MA_Mult) en combinant cinq modèles acoustiques monolingues (MA_Mono). Ces modèles monolingues sont ceux des langues anglaise (EN), française (FR), vietnamienne (VN), khmère (KH) et chinoise (CH). La méthode de combinaison « *ML-Sep* » [Schultz, 2001] est utilisée pour combiner ces modèles acoustiques monolingues en plusieurs modèles acoustiques multilingues. Le tableau C.1 présente les caractéristiques des modèles acoustiques des cinq langues que nous avons utilisées pour créer les modèles acoustiques multilingues.

Comme nous ne savons pas quels modèles acoustiques multilingues seront les meilleurs dans la tâche de transcription automatique du mo piu, nous nous proposons de comparer les résultats d'analyse des séquences phonétiques de trois différents modèles acoustiques multilingues.

- MA_Mult_VN-CH : comme la minorité mo piu se situe géographiquement aux frontières du Viêtnam et de la Chine, nous essayons de transcrire la langue mo piu en

utilisant les unités acoustiques et phonétiques du vietnamien (VN) et du chinois mandarin (CH).

- MA_Mult_VN-CH-KH : nous ajoutons un autre modèle acoustique asiatique disponible ; celui de la langue khmère (KH).
- MA_Mult_VN-CH-KH-EN-FR : cette fois-ci, nous ajoutons deux langues occidentales (français et anglais) pour proposer un modèle acoustique multilingue plus étendu.

Si nous considérons le premier modèle bilingue MA_Mult_VN-CH comme le modèle multilingue « minimal », nos objectifs sont : 1) de regarder si le fait d'ajouter une langue de la région (langue asiatique) en utilisant le modèle trilingue MA_Mult_VN-CH-KH peut améliorer la performance du système de transcription, 2) de déterminer si l'utilisation de plusieurs langues de familles différentes (langues asiatiques et langues occidentales) en utilisant le modèle MA_Mult_VN-CH-KH-EN-FR, et le fait d'avoir une couverture acoustique plus grande, peut s'avérer une solution efficace pour la transcription automatique du mo piu.

| **Nom du modèle acoustique** | **Quantité de données d'entraînement** | **Nombre des unités acoustiques** | **Autres caractéristiques** |
|---|---|---|---|
| BREF120 (français, FR) [Lamel, 1991] | > 100h | 43 | - Parole lue<br>- HMM de 3 états<br>- Nombre de gaussiennes : 16<br>- Contexte Indépendant : CI |
| WSJ (anglais, EN) [Paul, 1992] | ~ 20h | 40 | |
| KhmerASRcorpus (khmer, KH) [Seng, 2008] | ~5h | 36 | |
| VNSpeechCorpus (vietnamien, VN) [Le, 2004] | ~ 16h | 41 | |
| CADCC (mandarin, CH) [CCC, 2005] | ~ 5h | 34 | |

*Tableau C.1 : Cinq modèles acoustiques monolingues utilisés pour créer MA_Mult*

### C.2.2. Correspondances entre phonème (*Phone mapping*) :

Un tableau de correspondance phonétique entre les phonèmes au format X-SAMPA et au format API a été créé (tableau C.1). Chaque phonème « X-SAMPA » est étiqueté par sa langue (ex : *e_CH* signifie le phonème /e/ du mandarin). Pour représenter tous les phonèmes des cinq langues (FR, EN, CH, KH et VN), nous avons ainsi créé un tableau à 194 entrées, proposant pour chaque entrée deux colonnes : la première colonne présente un phonème au format X-SAMPA et la deuxième son équivalent en format API.

## C.3. Évaluation

### C.3.1. Protocole d'évaluation

Nous avons demandé à deux experts en phonétique et à des chercheurs en traitement de la parole ayant une bonne connaissance en phonétique d'évaluer les résultats des systèmes de transcription automatique fondés sur les trois modèles acoustiques bilingue, trilingue et multilingue (MA_Mult_VN-CH, MA_Mult_VN-CH-KH, MA_Mult_VN-CH-KH-EN-FR).

Après avoir écouté le son et observé le phonème généré par le système, nous avons demandé aux experts de lui donner un score parmi les quatre évaluations proposées : bon (b), voisin (v), acoustique (a) et faux (f).

- « bon » signifie que le phonème transcrit correspond à ce qui est entendu.
- « voisin » signifie que la transcription proposée n'est pas exacte, mais qu'elle est acoustiquement proche (les phonèmes appartiennent à une même classe articulatoire, ou ont le même lieu d'articulation). Par exemple, les consonnes /p/, /t/ et /k/ sont considérées comme des phonèmes voisins (consonnes plosives sourdes), et de même les consonnes /p / et /b/ sont aussi considérées comme voisines (même lieu articulatoire).
- « acoustique » signifie que l'unité transcrite n'est pas fausse, mais qu'elle porte sur une section de l'unité phonétique seulement.
- « faux » signifie qu'à l'évidence, la transcription proposée ne correspond en rien au son considéré.

Dans notre évaluation préliminaire de la transcription phonétique, neuf fichiers de parole du mo piu sont ainsi traités par nos trois différents systèmes acoustico-phonétiques multilingues ; on a donc évalué 27 fichiers correspondant à environ 742 événements

vocaliques. Les événements vocaliques sont les unités phonétiques (les consonnes et les voyelles) et les pauses.

La figure C.2 illustre un exemple d'évaluation de la transcription automatique d'une phrase en mo piu (l'outil « Praat » [Boersma, 2001] a été utilisé pour faciliter notre évaluation).

Dans la figure C.2, il y a deux points à considérer.

- L'une des caractéristiques particulières du mo piu est que cette langue posséde des voyelles qui semblent particulièrement longues. Les locuteurs parlent un peu comme s'ils chantaient. Une conséquence de cette longueur est une extinction lente de la phonation, ce qui peut générer un changement de voyelle (non souhaité) dans l'annotation, ou bien une fin de voyelle considérée comme aspirée ou nasalisée. Par exemple, dans la figure C.2, la transcription automatique donne les phonèmes /ɤ/+/ŋ/ alors qu'à l'écoute il semble que seule la voyelle (phonème /ɤ/) soit prononcée (voir le 5$^{ème}$ phonème de la référence dans la figure).
- L'évaluation doit aussi prendre en compte des erreurs sur les frontières des événements vocaliques, produites par les systèmes automatiques (les frontières des phonèmes sont représentées par les lignes en bleu dans la figure C.2).

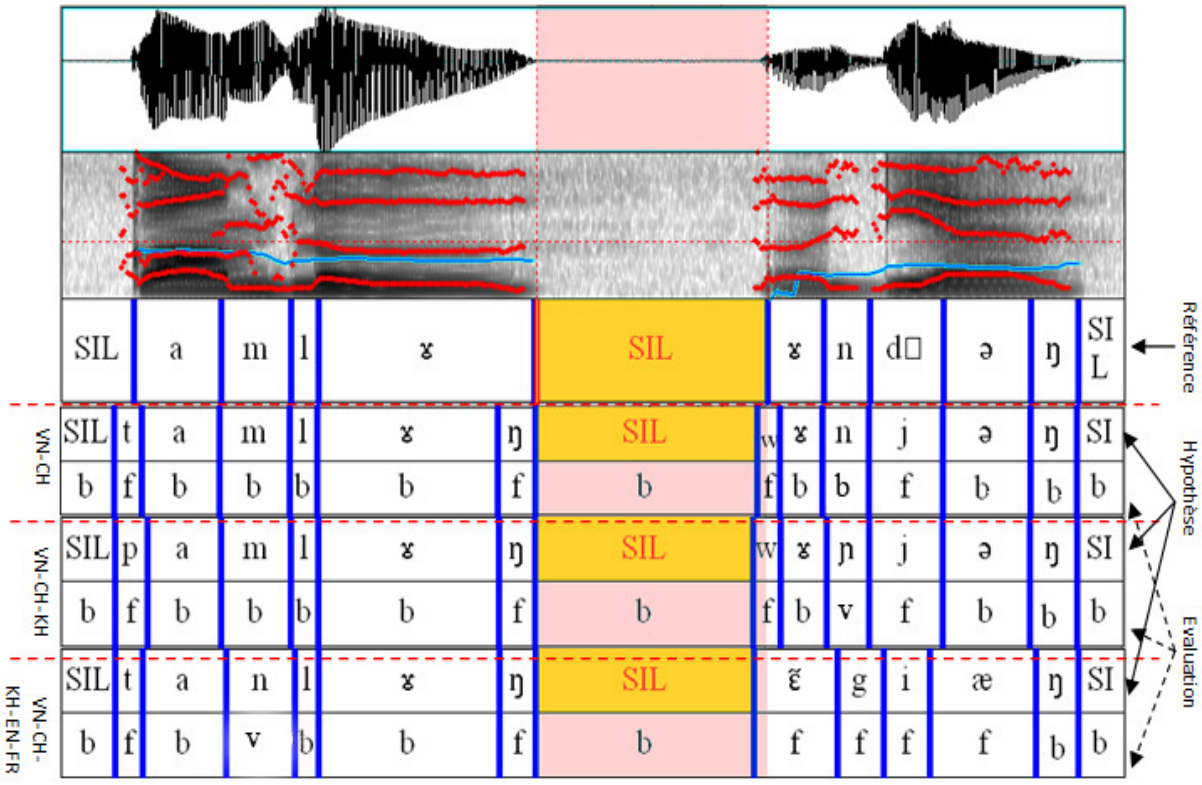



*Figure C.2 : Exemple d'évaluation de la transcription automatique d'une phrase prononcée en mo piu*

### C.3.2. Résultats de cette évaluation préliminaire

Nous mesurons trois taux d'erreur pour les erreurs commises par les systèmes de transcription automatique :

- taux d'erreur sur les frontières des événements vocaliques ;
- taux d'erreur sur les voyelles ;
- taux d'erreur sur les consonnes.

La figure C.3 illustre les taux d'erreur sur les frontières de tous les événements vocaliques produits par les trois systèmes acoustico-phonétiques multilingues. La figure C.4 illustre la répartition des taux d'erreurs sur les frontières en fonction de la durée, pour chaque système de transcription phonétique. Ces erreurs sont été déterminées par des experts en phonétique.

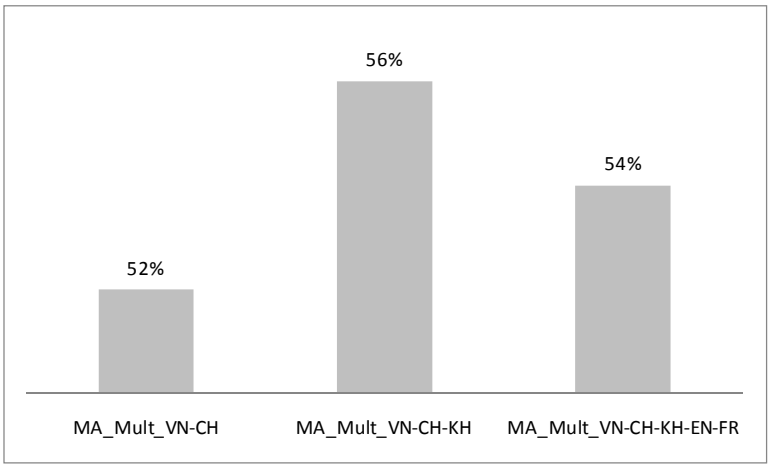



*Figure C.3 : Comparaison des taux d'erreur sur les frontières entre les différents systèmes acoustico-phonétiques multilingues*

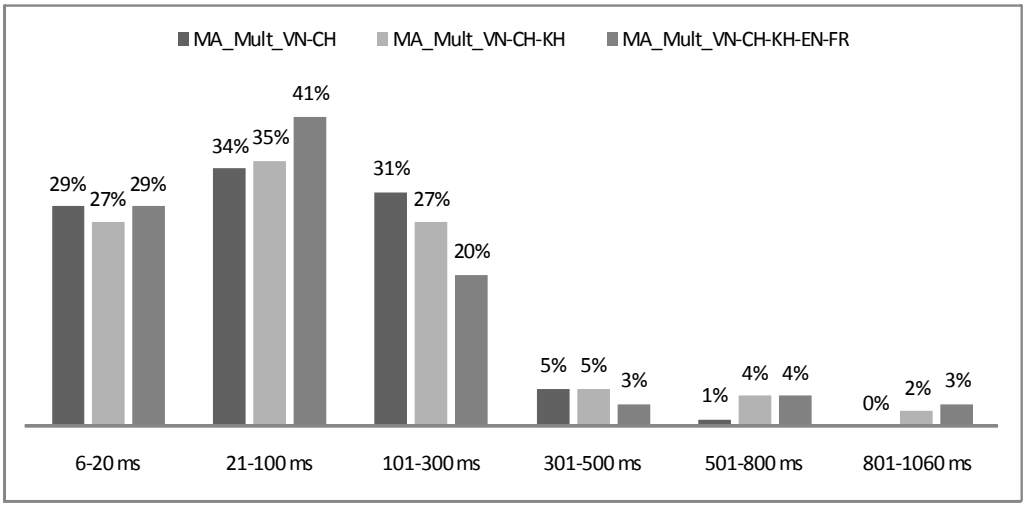



*Figure C.4 : Répartition des taux d'erreur sur les frontières en fonction de la durée des événements vocaliques*

Dans la figure C.3, nous observons que les différences des taux d'erreur sur les frontières d'un système de transcription automatique à un autre ne sont qu'au maximum de 4 %. De plus, les taux d'erreur sur les frontières illustrés dans la figure C.4 montrent que ceux-ci sont plus importants pour les événements vocaliques d'une durée inférieure à 300 ms. Au cours de sa première étude sur le mo piu, [Caelen-Haumont, 2010] a montré que les voyelles du mo piu sont généralement longues (de 110 ms à 1000 ms, avec une durée moyenne de 591 ms). Il semble donc que, dans cette étude préliminaire, les taux d'erreur sur les frontières des consonnes (qui sont des événements vocaliques de courte durée, comparés aux voyelles) sont plus importants que ceux des voyelles.

Si nous considérons maintenant l'analyse phonétique sous l'angle d'une comparaison entre les performances humaines et automatiques, alors, parmi les 742 événements vocaliques, il y a 117 voyelles et 128 consonnes qui sont elles aussi évaluées selon la grille : bon (b), voisin (v), acoustique (a) et faux (f).

Les figures C.5 et C.6 illustrent les évaluations des 117 voyelles pour l'ensemble des résultats produits par les trois systèmes automatiques, et séparément pour chacun des trois systèmes.

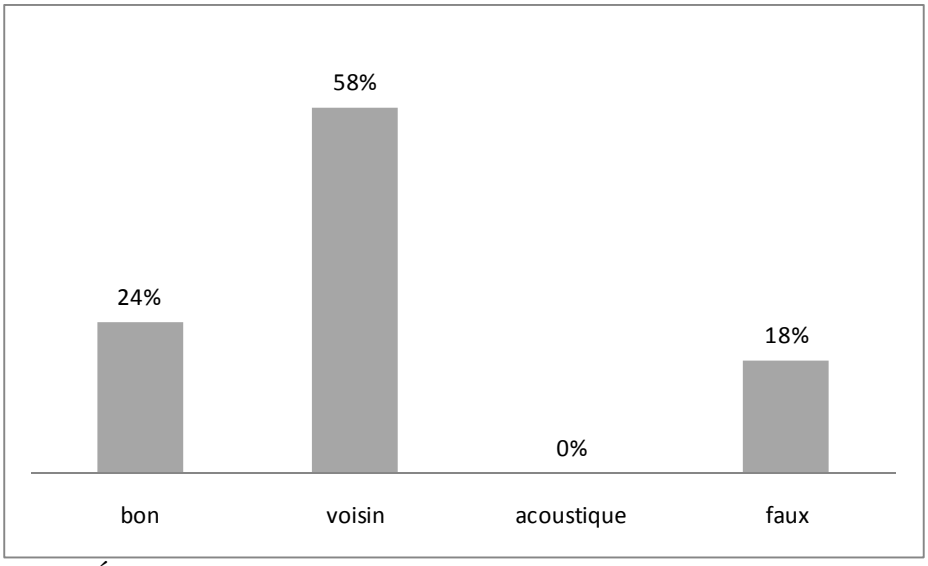



*Figure C.5 : Évaluation de toutes les voyelles pour l'ensemble des résultats des trois systèmes de transcription automatique*

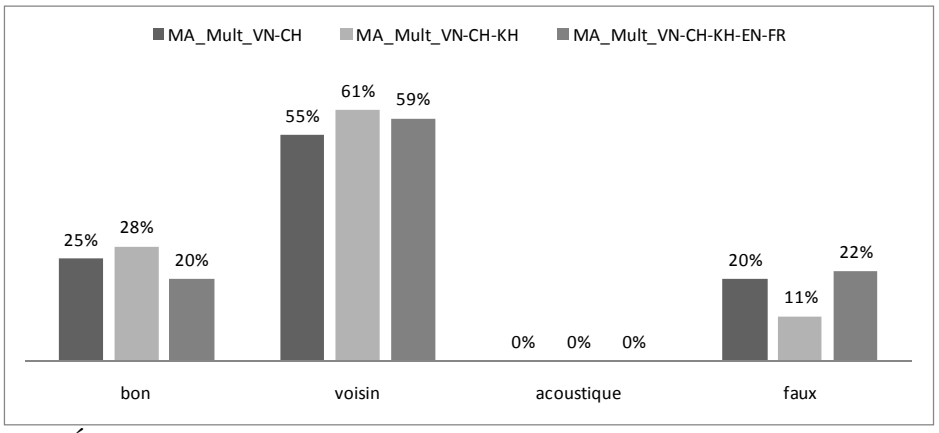



*Figure C.6 : Évaluation des erreurs de toutes les voyelles pour chacun des trois systèmes de transcription automatique*

Selon les résultats d'évaluation illustrés par la figure C.5, les systèmes automatiques fournissent 82 % des voyelles correctes ou phonétiquement voisines. Plus précisément, si nous comparons les résultats d'évaluation des trois systèmes de transcription séparément (figure C.6), les systèmes de transcription automatique fondés sur les deux modèles acoustiques bilingue (MA_Mult_VN-CH) et trilingue (MA_Mult_VN-CH-KH), qui contiennent la description de langues géographiquement proches du mo piu (langues asiatiques), présentent des résultats meilleurs que le système fondé sur le modèle acoustique quintilingue (MA_Mult_VN-CH-KH-EN-FR) dans lequel deux des cinq langues sont des langues occidentales (l'anglais et le français).

La figure C.7 résume les évaluations sur l'ensemble des 128 consonnes pour les résultats de transcription des trois systèmes réunis, et la figure C.8 illustre ces mêmes évaluations des consonnes, mais pour chaque système de transcription considéré séparément.

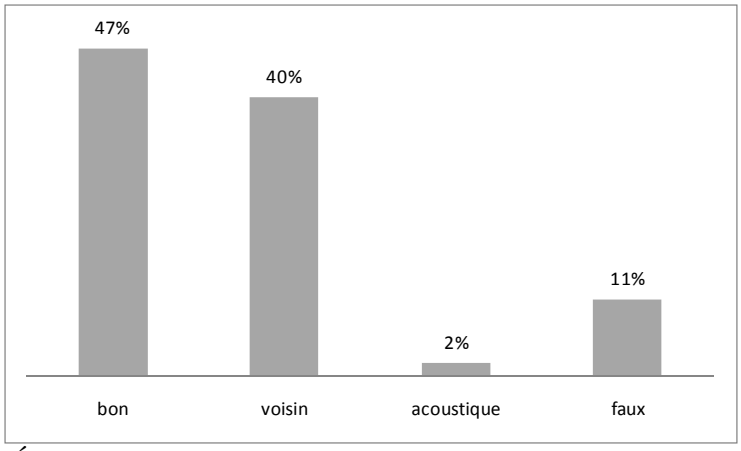



*Figure C.7 : Évaluation des 128 consonnes pour les résultats des trois systèmes de transcription automatique réunis*

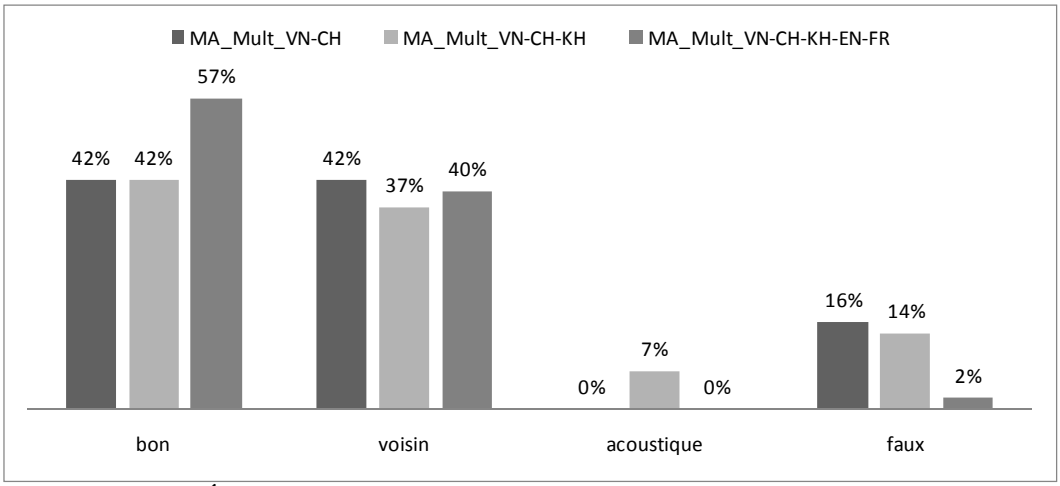



*Figure C.8 : Évaluation des 128 consonnes pour chacun des trois systèmes de transcription automatique*

Nous observons que 89 % des consonnes fournies par les trois systèmes automatiques sont phonétiquement exactes, ou phonétiquement ou acoustiquement proches (figure C.7). Par contre (figure C.8), et contrairement au cas des voyelles, pour l'évaluation de consonnes, le système de transcription automatique basé sur le modèle acoustique multilingue de cinq langues (MA_Mult_VN-CH-KH-EN-FR) semble meilleur que ceux fondés sur les autres modèles acoustiques bilingue et trilingue. La raison en est que le système quintilingue détecte mieux les nasales fréquentes en mo piu.

Toutefois, il est important de préciser que ces évaluations sont **préliminaires** et on été faites sur seulement 27 fichiers, qui représentent seulement 742 événements vocaliques. Pour cette raison, les résultats présentés ne sont pas très exhaustifs.

Cependant, nous pouvons tirer de cette étude une première conclusion simple : l'utilisation de systèmes de transcription automatique basés sur des modèles acoustiques multilingues semble être utile pour l'analyse de langues inconnues et qui n'ont pas de ressources écrites, comme le mo piu. Ces outils automatiques peuvent aider les phonéticiens à réaliser les analyses acoustiques, phonologiques et phonétiques de langues en danger comme le mo piu plus rapidement et à plus grande d'échelle, car ils permettent d'automatiser une grande partie des tâches de transcription. Cependant, il est clair que l'utilisation de tels systèmes ne peut se faire « seule » et que les résultats doivent être validés par des experts linguistiques, car les systèmes automatiques produisent des erreurs qui peuvent s'avérer importantes dans certain cas.

Ce travail débouche sur une deuxième phase, à savoir l'amélioration des résultats. À l'heure actuelle, d'autres études sont menées au sein du laboratoire pour déterminer si un autre

groupe de langues pourrait être plus efficace, à savoir : VN + CH + KH + FR. Les premiers résultats devraient être obtenus à l'automne 2011.

# Annexe D :
# Publications personnelles

## Conférences internationales


[Sam, 2011] **Sam, S.**, Xiao, X., Besacier, L., Castelli, E., Li, H., & Chng, E. S. (2011). *Speech Modulation Features for Robust Nonnative Speech Accent Detection.* Proc. Interspeech, Florence, Italy.

[Sam, 2010b] **Sam, S.**, Besacier, L., Castelli, E., Ma, B., Leung, C., & Li, H. (2010). *Autonomous acoustic model adaptation for multilingual meeting transcription involving high- and low-resourced languages*. Proc. SLTU, pp. 116-121, Penang, Malaysia.

[Sam, 2010c] **Sam, S.**, Castelli, E., & Besacier, L. (2010). *Unsupervised acoustic model adaptation for multi-origin non native.* Proc. Interspeech, pp. 254-257. Makuhari, Japan.

[Seng, 2008a] Seng, S., **Sam, S.**, Besacier, L., Bigi, B., & Castelli, E. (2008). *First Broadcast News Transcription System for Khmer Language*. Proc. LREC, pp. 2658-2661, Marrakech, Morocco.

[Seng, 2008b] Seng, S., **Sam, S.**, Le, V., Besacier, L., & Bigi, B. (2008). *Which units for acoustic and language modelling for khmer automatic speech recognition?* Proc. SLTU, pp. 33–38, Hanoi, Vietnam.


## Conférences francophones


[Caelen-Haumont, 2011] Caelen-Haumont, G., Hai, B. P., Trang, D. D., Salmon, J.-P., & **Sam, S.** (2011). *Démonstration du logiciel d'analyse de la parole Praat-MOMEL-MELISM, présentation et écoute du corpus mo piu*. Proc. CERLICO, Orléans, France.

[Sam, 2010a] **Sam, S.**, Besacier, L., & Castelli, E. (2010). *Adaptation autonome de modèles acoustiques pour la transcription automatique de réunions multilingues*. Proc. JEP(Journées d'Études sur la Parole), Mons, Belgique.

[Sam, 2009] **Sam, S.** (2009). *Vers des modèles autonomes pour la reconnaissance automatique de la parole multilingue*. RJCP (Rencontre des Jeunes Chercheurs en Parole), Avignon, France.

# Bibliographie


[Adami, 2003] Adami, A., & Hermansky, H. (2003). *Segmentation of speech for speaker and language recognition*, Eurospeech, pp. 841-844, Geneva, Switzerland.

[Adda-Decker, 2003] Adda-Decker, M., Antoine, F., de Mareuil, P., Vasilescu, I., Lamel, L., Vaissiere, J., et al. (2003). *Phonetic knowledge, phonotactics and perceptual validation for automatic language identification*, Proc. ICPhS, pp. 747–750, Barcelona, Spain.

[Ahn, 2000] Ahn, S., Kang, S., & Ko, H. (2000). *Effective speaker adaptations for speaker verification*. ICASSP, 1081–1084, Istanbul,Turkey.

[Aliprand, 2003] Aliprand, J., Allen, J., Becker, J., Davis, M., Everson, M., Freytag, A., et al. (2003). The Unicode standard, version 4.0 (Vol. 45): Addison-Wesley Professional, 1504 p., ISBN : 0321185781.

[Arslan, 1996] Arslan, L., & Hansen, J. (1996). Language Accent Classification in American English. *Speech Communication, vol.18*, pp 353–367.

[Atal, 1979] Atal, B., & Schroeder, M. (1979). Predictive coding of speech signals and subjective error criteria. *IEEE Transactions on Acoustics, Speech and Signal Processing, vol.27* (3), pp. 247-254.

[Baker, 1975] Baker, J. (1975). Stochastic Modeling for Automatic Speech Recognition. *Speech Recognition*, Academic Press, Academic Press, pp. 521-542.

[Barras, 2001] Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production* 1. *Speech Communication, vol.33* (1-2), pp. 5-22.

[Baum, 1970] Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics, vol.41* (1), pp. 164-171.

[Berment, 2004] Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues «peu dotées».* Thèse de Doctorat, 277 p., UJF, Grenoble, France.

[Billa, 2002] Billa, J., Noamany, M., Srivastava, A., Liu, D., Stone, R., Xu, J., et al. (2002). *Audio indexing of Arabic broadcast news*, Acoustics, Speech, and Signal Processing, IEEE International Conference, pp. 5-8.

[Bisani, 2003] Bisani, M., & Ney, H. (2003). *Multigram-based grapheme-to-phoneme conversion for LVCSR*, Eurospeech, pp. 933-936, Geneva, Switzerland..

[Boersma, 2001] Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot international, vol.5* (9/10), pp. 341-345.

[Burges, 1998] Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery, vol.2* (2), pp. 121-167.

[Caelen-Haumont, 2010] Caelen-Haumont, G., et al. (2010). *Mo Piu minority language: data base, first steps and first experiments.* Proc. SLTU, pp. 42-50, Penang, Malaysia.

[Caelen-Haumont, 2011] Caelen-Haumont, G., Hai, B. P., Trang, D. D., Salmon, J.-P., & **Sam, S.** (2011). *Démonstration du logiciel d'analyse de la parole Praat-MOMEL-MELISM, présentation et écoute du corpus mo piu*. Proc. CERLICO, Orléans, France.

[CCC, 2005] CCC. (2005). CCC resources : Online Chinese Corpus Consrtium Retrieved from http://www.dear.com/CCC/resources.htm

[Dai, 2003] Dai, P., Iurgel, U., & Rigoll, G. (2003). *A novel feature combination approach for spoken document classification with support vector machines*, Proc. Multimedia Information Retrieval Workshop, pp. 1-5, Toronto, Canada.

[Dempster, 1977] Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), vol.39* (1), pp. 1-38.

[Duda, 2000] Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification* (Vol. 2): Wiley-Interscience, 654 p., ISBN : 0471056693.

[Flege, 1995] Flege, J. (1995). Second language speech learning: Theory, findings, and problems. *In W. Strange (Ed.), Speech perception and linguistic experience: Issues in cross-language research*, pp. 233-277.

[Flege, 1997] Flege, J., Frieda, E., & Nozawa, T. (1997). Amount of native language (L1) use affects the pronunciation of an L2. *Journal of Phonetics, vol.25*, pp. 169-186.

[Flege, 2004] Flege, J., & MacKay, I. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition, vol.26* (01), pp. 1-34.

[Gales, 1996] Gales, M., & Woodland, P. (1996). Mean and variance adaptation within the MLLR framework. *Computer speech and language, vol.10* (4), pp. 249-264.

[Gauvain, 1994] Gauvain, J., & Lee, C. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing, vol.2* (2), pp. 291-298.

[Gauvain, 2000] Gauvain, J., & Lamel, L. (2000). Large-vocabulary continuous speech recognition: advances and applications. *Proceedings of the IEEE, vol.88* (8), pp. 1181-1200.

[Gauvain, 2002] Gauvain, J., Lamel, L., & Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech Communication, vol.37* (1-2), pp. 89-108.

[Gonzalez, 2002] Gonzalez, R., & Woods, R. (2002). *Digital image processing*: Prentice Hall, 793 p., ISBN: 0201180758.

[Graddol, 1997] Graddol, D. (1997). The future of English? A guide to forecasting the popularity of the English language in the 21st century. *The British Concil (www.britishcouncil.org/learning-elt-future.pdf)*.

[Grover, 1987] Grover, C., Jamieson, D., & Dobrovolsky, M. (1987). Intonation in English, French and German: perception and production. *Language and Speech, vol.30* (3), SAGE Publications, pp. 277-295.

[Hansen, 1995] Hansen, J. H. L., & Arslan, L. M. (1995). *Foreign accent classification using source generator based prosodic features.* Proc. ICASSP, pp. 836-839, Michigan, USA.

[Haton, 1991] Haton, J., Pierrel, J., Pérennou, G., Caelen, J., & Gauvain, J. (1991). *Reconnaissance automatique de la parole*: Dunod, 239 p., ISBN : 2040188274.

[Hazen, 1993] Hazen, T. J., & Zue, V. W. (1993). *Automatic language identification using a segment-based approach.* Proc. Eurospeech, Cambridge, pp. 1307-1310, Berlin, Germany.

[Hermansky, 1991] Hermansky, H., & Cox Jr, L. (1991). *Perceptual Linear Predictive (PLP) analysis-resynthesis technique*, Applications of Signal Processing to Audio and Acoustics. Final Program and Paper Summaries., IEEE ASSP Workshop, pp. 037-038.

[Hieronymus, 1993] Hieronymus, J. (1993). ASCII Phonetic Symbols for the World's Languages: Worldbet. *Journal of the Internation Phonetic Association, vol.23*.

[Huang, 2001] Huang, X., Acero, A., & Hon, H. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*: Prentice Hall PTR Upper Saddle River, NJ, USA, 1008 p., ISBN : 0130226165.

[IPA, 1999] IPA. (1999). *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*: Cambridge: Cambridge University Press, 214 p., ISBN : 0521637511.

[Jain, 2001] Jain, P., & Hermansky, H. (2001). *Improved mean and variance normalization for robust speech recognition.* Proc. ICASSP, pp. 4015-4015, Salt Lake City, Utah.

[Jelinek, 1976] Jelinek, F. (1976). Continous speech recognition by statistical methods. *Proceedings of the IEEE, vol.64:4,* pp. 532-556.

[Joachims, 2002] Joachims, T. (2002). Learning to classify text using support vector machines: Methods, theory, and algorithms. *Computational Linguistics, vol.29* (4), pp. 656-664.

[Kanedera, 1999] Kanedera, N., Arai, T., Hermansky, H., & Pavel, M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication, vol.28* (1), pp. 43-55.

[Kim, 1997] Kim, K., Relkin, N., Lee, K., & Hirsch, J. (1997). Distinct cortical areas associated with native and second languages. *Nature, vol.388* (6638), pp. 171-174.

[Kinnunen, 2006] Kinnunen, T. (2006). *Joint acoustic-modulation frequency for speaker recognition.* Proc. ICASSP, pp. 14-19, Toulouse,France.

[Kirchhoff, 2002] Kirchhoff, K., Parandekar, S., & Bilmes, J. (2002). *Mixed-memory Markov models for automatic language identification*, Proc. ICASSP, pp. 761-764, Orlando, Florida, USA.

[Kneser, 1995] Kneser, R., & Ney, H. (1995). *Improved backing-off for m-gram language modeling.* Proc. ICASSP, pp. 181-184, Detroit, Michigan, USA.

[Kuhl, 2000] Kuhl, P. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences of the United States of America, vol.97* (22), pp. 11850-11857.

[Lamel, 1991] Lamel, L., Gauvain, J., & Eskenazi, M. (1991). *BREF, a large vocabulary spoken corpus for French*, Proc. Eurospeech, pp. 24-26, Genove, Italy.

[Lazzari, 2006] Lazzari, G., & Steinbiss, V. (2006). Human Language Technologies for Europe. *ITC IRST/TC-Star project report*.

[Lazzarotto, 2007] Lazzarotto, R. (2007). *Réalisation d'un système de communication multimédia SIAM pour espaces perceptifs (convergence entre appareils communicants, sécurité, intégration de modules)*, Master thesis, ESIL, Marseille, France.

[Le, 2004] Le, V., Tran, D., Castelli, E., Besacier, L., & Serignat, J. (2004). *Spoken and written language resources for Vietnamese*, Proc. LREC, pp. 599–602, Lisbon, Portugal.

[Le, 2006] Le, V. (2006). *Reconnaissance automatique de la parole pour des langues peu dotées.* Thèse de doctorat, Université Joseph Fourier-Grenoble 1, 200 p., Grenoble, France.

[Lee, 2002] Lee, K., Hon, H., & Reddy, R. (2002). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing, vol.38* (1), pp. 35-45.

[Leggetter, 1995] Leggetter, C., & Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer speech and language, vol.9* (2), pp. 171-185.

[L'Expansion, 2007] L'Expansion. (2007). Le nombre de touristes étrangers en France en 2006, *L'Expansion (http://lexpansion.lexpress.fr), Paris*.

[Li, 2007] Li, H., Ma, B., & Lee, C. (2007). A vector space modeling approach to spoken language identification. *IEEE Transactions on Audio, Speech and Language Processing, vol.15* (1), pp. 271-284.

[Liu, 2006] Liu, Y., & Fung, P. (2006). *Multi-accent chinese speech recognition.* ICSLP, pp. 133-136, Pittsburgh, USA.

[Livescu, 1999] Livescu, K. (1999). *Analysis and modeling of non-native speech for automatic speech recognition*, Master thesis, 89 p., MIT-department of Electrical Engineering and Computer Science.

[Lyer, 1997] Lyer, R., Ostendorf, M., & Meteer, M. (1997). *Analyzing and predicting language model improvements*, Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 254-261, New York, USA.

[Ma, 2002] Ma, B., Guan, C., Li, H., & Lee, C. (2002). *Multilingual speech recognition with language identification*, Proc. 7th International Conference on Spoken Language Processing, pp. 505-508, Colorado, USA.

[Ma, 2005] Ma, B., & Li, H. (2005). *A phonotactic-semantic paradigm for automatic spoken document classification*, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 369-376, Salvador, Brazil.

[Markel, 1982] Markel, J., & Gray, A. (1982). *Linear prediction of speech*: Springer-Verlag, New York, USA, 288 p., ISBN : 3540075631.

[Martin, 1997a] Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). *The DET curve in assessment of detection task performance.*

Proc. European Conference on Speech Communication and Technology, pp. 1895-1898, Rhodes, Greece.

[Martin, 1997b] Martin, S., Liermann, J., & Ney, H. (1997). *Adaptive topic-dependent language modelling using word-based varigrams*, In Proc. Eurospeech'97, pp. 1447-1450.

[Matrouf, 1998] Matrouf, D., Adda-Decker, M., Lamel, L., & Gauvain, J. (1998). *Language identification incorporating lexical information*, Fifth International Conference on Spoken Language Processing, pp. 181-184.

[McDonough, 1997] McDonough, J., Anastasakos, T., Zavaliagkos, G., & Gish, H. (1997). *Speaker-adapted training on the switchboard corpus.* Proc. ICASSP, pp. 1059-1062, Munich, Germany.

[Nagarajan, 2004] Nagarajan, T., & Murthy, H. (2004). *Language identification using parallel syllable-like unit recognition*, ICASSP, pp. 401-404, Montreal, CA.

[Nazari, 2008] Nazari, M., Sayadiyan, A., & Valiollahzadeh, S. (2008). Probabilistic SVM/GMM Classifier for Speaker-Independent Vowel Recognition in Continues Speech. *ACM Computing Research Repository, vol. abs/0812.2411*.

[O'Grady, 2000] O'Grady, W., & Archibald, J. (2000). *Contemporary Linguistic Analysis: An Introduction*: Addison Wesley Publishing Company, 688 p., ISBN : 0201478129.

[Oh, 2007] Oh, Y., Yoon, J., & Kim, H. (2007). Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication, vol.49* (1), pp. 59-70.

[Pallett, 2003] Pallett, D. (2003). *A look at NIST'S benchmark ASR tests: past, present, and future.* Proc. ASRU, pp. 483-488, Virgin Islands, USA.

[Paul, 1992] Paul, D., & Baker, J. (1992). *The design for the Wall Street Journal-based CSR corpus*, Proc. of DARPA Workshop on SNL, pp. 357-362, San Mateo, CA.

[Piat, 2008] Piat, M., Fohr, D., & Illina, I. (2008). *Foreign accent identification based on prosodic parameters.* Proc. Interspeech, pp. 759-762, Brisbane, Australia.

[Quénot, 2010] Quénot, G., Tan, T., Le, V., Ayache, S., Besacier, L., & Mulhem, P. (2010). Content-based search in multilingual audiovisual documents using the International Phonetic Alphabet. *Multimedia Tools and Applications, vol.48* (1), pp. 123-140.

[Rabiner, 1993] Rabiner, L., & Juang, B. (1993). *Fundamentals of speech recognition*: Prentice hall Englewood Cliffs, New Jersey, 496 p., ISBN : 0130151572.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE, vol.77* (2), pp. 257-286.

[Rochet, 1995] Rochet, B. (1995). Perception and production of second-language speech sounds by adults. *Speech perception and linguistic experience: Issues in cross-language research*, York Press, pp. 379-410.

[Rosen, 1991] Rosen, S., & Howell, P. (1991). *Signals and systems for speech and hearing*: Academic Press, 322 p., ISBN : 0125972318.

[Sam, 2009] [Sam, 2009] **Sam, S.** (2009). *Vers des modèles autonomes pour la reconnaissance automatique de la parole multilingue*. RJCP (Rencontre des Jeunes Chercheurs en Parole), Avignon, France.

[Sam, 2010a] **Sam, S.**, Besacier, L., & Castelli, E. (2010). *Adaptation autonome de modèles acoustiques pour la transcription automatique de réunions multilingues*. Proc. JEP (Journées d'Études sur la Parole), Mons, Belgique.

[Sam, 2010b] **Sam, S.**, Besacier, L., Castelli, E., Ma, B., Leung, C., & Li, H. (2010). *Autonomous acoustic model adaptation for multilingual meeting transcription involving high- and low-resourced languages*. Proc. SLTU, pp. 116-121, Penang, Malaysia.

[Sam, 2010c] **Sam, S.**, Castelli, E., & Besacier, L. (2010). *Unsupervised acoustic model adaptation for multi-origin non native*. Proc. Interspeech, pp. 254-257. Makuhari, Japan.

[Sam, 2011] Sam, S., Xiao, X., Besacier, L., Castelli, E., Li, H., & Chng, E. S. (2011). *Speech Modulation Features for Robust Nonnative Speech Accent Detection*. Interspeech 2011, Florence, Italy.

[Samudravijaya, 2003] Samudravijaya, K., & Barot, M. (2003). *A comparison of public-domain software tools for speech recognition.* Proc. WSLP, pp. 125-131, Mumbai, India.

[Sankoff, 1999] Sankoff, D., & Kruskal, J. (1999). *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison* (second edition ed.): CSLI, 408 p., ISBN : 1575862174.

[Schultz, 2000] Schultz, T., & Waibel, A. (2000). *Polyphone decision tree specialization for language adaptation*, Proc. ICASSP, pp. 1707-1710, Istanbul,Turkey.

[Schultz, 2001] Schultz, T., & Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication, vol.35* (1), pp. 31-52.

[Schultz, 2002] Schultz, T. (2002). *Globalphone: a multilingual speech and text database developed at Karlsruhe University.* ICSLP, pp. 345-348, Colorado, USA.

[Schultz, 2006] Schultz, T., & Kirchhoff, K. (2006). *Multilingual speech processing*: Academic Press, 536 p., ISBN : 0120885018

[Schwenk, 2002] Schwenk, H., & Gauvain, J. (2002). *Connectionist language modeling for large vocabulary continuous speech recognition.* Proc. ICASSP, pp. 765-768, Florida, USA.

[Schwenk, 2007] Schwenk, H. (2007). Continuous space language models. *Computer Speech & Language, vol.21* (3), pp. 492-518.

[Seng, 2008a] Seng, S., **Sam, S.**, Besacier, L., Bigi, B., & Castelli, E. (2008). *First Broadcast News Transcription System for Khmer Language*. Proc. LREC, pp. 2658-2661, Marrakech, Morocco.

[Seng, 2008b] Seng, S., **Sam, S.**, Le, V., Besacier, L., & Bigi, B. (2008). *Which units for acoustic and language modelling for khmer automatic speech recognition?* Proc. SLTU, pp. 33–38, Hanoi, Vietnam.

[Shannon, 1949] Shannon, C. (1949). Communication in the presence of noise. *Proceedings of the IRE, vol.37* (1), pp. 10-21.

[Singer, 2003] Singer, E., Torres-Carrasquillo, P., Gleason, T., Campbell, W., & Reynolds, D. (2003). *Acoustic, phonetic and discriminative approaches to automatic language recognition*, Proc. Eurospeech, pp. 1345–1348, Geneva, Switzerland.

[Stolcke, 2002] Stolcke, A. (2002). *SRILM - an extensible language modeling toolkit*, ICSLP, pp. 901–904, Colorado, USA.

[Sugiyama, 1991] Sugiyama, M. (1991). *Automatic language recognition using acoustic features*, Proc. ICASSP, pp. 813-816, Ontario, CA.

[Tan, 2006] Tan, T.-P., & Besacier, L. (2006). *A French non-native corpus for automatic speech recognition.* Proc. LREC, pp. 1610-1613, Genoa, Italy.

[Tan, 2007] Tan, T., & Besacier, L. (2007). *Modeling Context and Language Variation for Non-Native Speech Recognition.* Proc. Interspeech, pp. 1429-1432, Antwerp, Belgium.

[Tan, 2008] Tan, T.-P. (2008). *Automatic Speech Recognition for Non-Native Speakers.* Thèse de doctorat, 175 p., UJF, Grenoble, France.

[Torres-Carrasquillo, 2002] Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., & Deller Jr, J. (2002). *Approaches to language identification using Gaussian mixture models and shifted delta cepstral features*, ICSLP, pp. 89-92, Colorado, USA..

[Touch, 2010] Touch, S., Besacier, L., Castelli, E., & Boitet, C. (2010). *Voice aided input for phrase selection using a low level ASR approach - Application to French and Khmer phrasebooks.* Proc. SLTU, pp. 111-115, Penang, Malaysia.

[Uebler, 1999] Uebler, U., & Boros, M. (1999). *Recognition of non-native German speech with multilingual recognizers*, Proc. Eurospeech, Volume 2, pp. 907-910.

[UNESCO, 2010] UNESCO. (2010). *Atlas of the World's Languages in Danger*: BERNAN PR, 218 p., ISBN : 9789231040962.

[Waibel, 2000] Waibel, A., Geutner, P., Tomokiyo, L., Schultz, T., & Woszczyna, M. (2000). Multilinguality in speech and spoken language systems. *Proceedings of the IEEE, vol.88* (8), pp. 1181-1190.

[Wanneroy, 1999] Wanneroy, R., Bilinski, C., Barras, C., Adda-Decker, M., & Geoffrois, E. (1999). *Acoustic-Phonetic Modeling of Non-Native Speech for Language Identification.* Proc. MIST,Leusden, The Netherlands.

[Wang, 2003] Wang, Z., Schultz, T., & Waibel, A. (2003). *Comparison of acoustic model adaptation techniques on non-native speech*, Proc. ICASSP, pp. 540-543, Hong Kong.

[Wells, 1995] Wells, J. (1995). Computer-coding the IPA: a proposed extension of SAMPA. *draft article*, University College London, England (http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm).

[Wells, 1997] Wells, J. (1997). SAMPA computer readable phonetic alphabet. *Handbook of Standards and Resources for Spoken Language Systems,* Mouton De Gruyter, pp. 684-730.

[White, 2008] White, C., Khudanpur, S., & Baker, J. (2008). *An Investigation of Acoustic Models for Multilingual Code-Switching.* Proc. Interspeech, pp. 2691-2694, Brisbane, Australia..

[Witt, 1999] Witt, S. (1999). *Use of speech recognition in computer-assisted language learning.* PhD. Thesis, 139 p., University of Cambridge.

[Wood, 1987] Wood, F., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometr. Intel. Lab. Syst, vol.2*, pp. 37–52.

[Xiong, 2009] Xiong, X. (2009). *Robust speech features and acoustic models for speech recognition.* PhD. Thesis, 194 p., Nanyang Technological University, Singapore.

[Young, 1999] Young, S. (1999). Acoustic modelling for large vocabulary continuous speech recognition. *NATO ASI Series F Computer and Systems Sciences, vol.169*, pp. 18-39.

[Young, 1994] Young, S., & Young, S. (1994). The HTK Hidden Markov Model toolkit: design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, Entropic Cambridge Research Laboratory, Ltd, pp. 2-44.

[Zissman, 1996a] Zissman, M. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing, vol.4* (1), pp. 31-44.

[Zissman, 1996b] Zissman, M. A., Gleason, T. P., Rekart, D. M., & Losiewicz, B. L. (1996). *Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech.* Proc. ICASSP, pp. 777-780, Atlenta, Georgia.

[Zue, 2000] Zue, V., & James, R. (2000). Conversational interfaces: Advances and challenges. *Proceedings of the IEEE, vol.88* (8), pp. 1166–1180.