Université Henri Poincaré
Master 2 Recherche, Reconnaissance Apprentissage Raisonnement
**Rapport de Stage Recherche**

Maîtres de stage : Dominique Fohr, Irina Illina et Denis Jouvet
Groupe de recherche : PAROLE

# Reconnaissance de parole avec prise de son distante pour la domotique

Luiza Orosanu

Laboratoire Lorrain de Recherche Informatique
615 r. du Jardin Botanique
54600 Villers-lès-Nancy



Nancy, le 22 juin 2011

# REMERCIEMENTS

Je tiens à remercier en premier lieu mes encadrants, M. Dominique Fohr, M. Denis Jouvet et MMe Irina Illina , qui m'ont permis de réaliser ce stage. Leurs compétences, leurs conseils et remarques ont été d'une aide précieuse tout au long de ce stage et plus généralement pour mon initiation à la recherche. J'ai pu également apprécier leur gentillesse et leur disponibilité malgré des emplois du temps chargés.

Je souhaite également remercier l'ensemble de l'équipe PAROLE pour m'avoir fourni un cadre de travail enrichissant et agréable.

Je tiens aussi à remercier M. Didier Galmiche qui m'a donné la chance de suivre ce Master et m'a donné les moyens de réaliser le stage et le rapport dans les meilleures conditions possibles.

Enfin, j'aimerais remercier mes sœurs, Andreea et Simona, et aussi mon ami, Luc, pour leur soutien et leur aide tout au long de ce stage.

**Résumé**

Actuellement la majorité des systèmes de reconnaissance de la parole fonctionnent dans un environnement calme avec une prise de son de bonne qualité (peu de bruit ambiant, locuteur proche du microphone, avec micro-casque par exemple). Selon l'application dans laquelle va intervenir le système de reconnaissance, il n'est pas toujours possible de contraindre l'utilisateur à porter un micro-casque ou à se rapprocher de l'appareil de prise de son, par exemple, dans un environnement domestique pour des applications telle que la domotique. Dans cette application en particulier, il est tout à fait inenvisageable d'imposer à l'utilisateur de telles contraintes peu importe les aménagements (microphone sans-fil). Typiquement, la prise de son, dans l'optique de cette application, devra se faire à partir de microphones prépositionnés dans une pièce, et, en conséquence, à distance de l'utilisateur, ce qui a pour effet de compliquer grandement le processus de reconnaissance.

Mon stage a consisté à étudier les performances d'un système de reconnaissance de la parole avec prise de son distante appliqué au domaine de la domotique. L'idée étant d'utiliser les connaissances du domaine de la reconnaissance et de les appliquer pour piloter l'équipement domotique d'une maison.

En plus de prendre en compte les problématiques usuelles du domaine de la reconnaissance de la parole, nous devions donc prendre en compte un certain nombre de difficultés sous-jacentes liées à la prise de son distante et à l'application au domaine de la domotique. Pour mettre une telle installation en place, nous devons considérer d'autres problèmes d'un ordre plus technique. Le système doit comprendre les instructions indépendamment du locuteur (il doit bien comprendre une instruction même si elle est prononcée par différents locuteurs ou par le même locuteur dans différents états émotionnels). Il doit aussi faire la différence entre les commandes et le langage naturel (résidents discutant entre eux, ...).

Ce stage a permis d'évaluer les performances et la faisabilité de la mise en place d'un tel système de reconnaissance. La qualité de la reconnaissance dépend fortement de la qualité du signal de parole à reconnaître et de l'adéquation entre les conditions d'utilisation (liées à l'environnement et ici à la prise de son à distance) et les conditions de collecte des données ayant servi à l'apprentissage des modèles acoustiques. Nous avons du créer une base de signaux collectés à distance (correspondants à une liste de commandes domotiques) que l'on a utilisé comme « entrée » du système de reconnaissance. Ces données peuvent néanmoins ne pas s'avérer suffisantes pour paramétrer et obtenir un système de reconnaissance efficace, nous les avons donc utilisés uniquement en tant que données de test.. Une des difficultés à laquelle nous avons été confrontés était l'essai d'un grand nombre de configurations, paramètres de reconnaissance et modèles de langage différents afin de déterminer la configuration conduisant à des performances optimales du système de reconnaissance.

# Table des matières

# Chapitre 1

# Introduction

Pour bien aborder la problématique de la « reconnaissance de parole avec prise de son distante pour la domotique », on commence par mentionner les principales définitions et problèmes qu'un tel concept peut poser. On termine le chapitre par les problèmes abordés pendant le stage, et les principales étapes du travail.

## 1.1 Reconnaissance de la parole

La communication orale représente une forme de relation entre les personnes. Elle est la manière la plus simple pour exprimer une opinion, une idée, un sentiment ou encore un désir. Ce moyen de communication, qui est le résultat de la volonté et de la pensée, est utilisé pour échanger, transmettre ou recevoir des informations.

La seule contrainte posée dans ce type d'interaction humaine est le partage d'une langue commune entre les interlocuteurs. Le but d'une conversation transparaît seulement si le message transmis est clair, correct, précis, ce qui permet au récepteur de le bien comprendre. Si le canal sonore n'est pas perturbé par des facteurs qui altèrent ou bloquent la communication (par exemple : le bruit ambiant), la réaction du récepteur sera la juste.

Aujourd'hui, la communication orale n'est pas utilisée uniquement entre les humains. En apprenant à une machine les connaissances acoustiques et linguistiques nécessaires pour bien comprendre la parole d'un humain, la communication entre l'homme et la machine dépendra essentiellement de la qualité du signal et de la performance des outils de reconnaissance. Les buts d'un tel apprentissage peuvent par exemple être de faciliter l'exécution de commandes, la traduction, la transcription, etc.

La reconnaissance de la parole est la technique qui permet l'analyse des sons captés par un microphone pour les transcrire sous forme d'une suite des mots exploitable par les machines [5].

L'utilisation de la parole comme mode de communication avec une machine apporte des avantages pour les utilisateurs non-specialisés, handicapés, ou étant dans l'indisponibilité d'utiliser leur mains. Un autre intérêt reste aussi dans la possibilité d'accéder aux machines à distance (par exemple, par téléphone).

Voici quelques exemples de domaines d'application de la reconnaissance automatique de la parole :

- machine à dicter et bureautique,
- systèmes d'identification de langues,
- traduction automatique,

– commandes de machines (« mains libres ») : téléphones mobiles, avions, hélicoptères, systèmes de guidage d'automobiles(GPS), maisons intelligentes, etc,
– systèmes d'intelligence ambiante ,
– jeux et jouets,
– support aux personnes handicapés.

## 1.2 Domaine de la domotique

*La domotique* est un ensemble de technologies liées à l'électronique, l'automatisme, l'informatique et les télécommunications mises en place dans un bâtiment pour faciliter la vie de ses résidents. Elle vise à apporter des fonctions de confort (gestion d'énergie, optimisation de l'éclairage et du chauffage), de sécurité (alarmes) et de communication (commandes à distance, téléphonie, ...), que l'on peut mettre en place dans les maisons, les hôtels, les lieux publics par exemple [19].

Voici quelques exemples d'outils qu'on peut contrôler dans un bâtiment équipé d'un tel système :

– les lumières
– les volets
– la porte
– le chauffage
– la climatisation
– le téléphone
– le téléviseur
– le système de sécurité
– l'arrosage automatique
– les lumières à l'extérieur

D'un point de vue plus pratique, la domotique est basée sur la mise en réseau des différents appareils électriques de la maison, contrôlés par une «intelligence» centralisée. L'intelligence qui gère ces commandes est une centrale programmable. Cette centrale est pilotée par les résidents pour réaliser les tâches souhaitées (la centrale peut aussi décider de prendre certaines décisions d'elle-même).

Typiquement, les outils de pilotage de la centrale sont des terminaux informatiques (ordinateurs de poche, smartphone, ...), des télécommandes ou encore une interface sur téléviseur. L'idée ici est d'utiliser la reconnaissance de la parole pour permettre de piloter la centrale par le biais de commandes vocales.

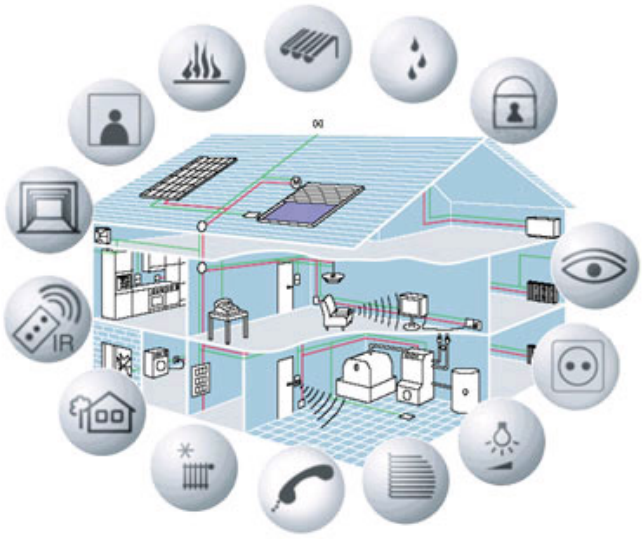



FIGURE 1.1 – Une maison équipée d'un système domotique.

## 1.3 Reconnaissance de la parole et la domotique

En plus de prendre en compte les problématiques usuelles lors de la mise en place d'un système de reconnaissance de la parole, nous allons devoir prendre en compte un certain nombre de difficultés sous-jacentes liées à la prise de son distante et à l'application au domaine de la domotique. Les principales difficultés auxquelles nous allons être confrontés sont :

- la qualité de l'apprentissage et l'adéquation des modèles (modèles statistiques)
  Comme mentionné avant, un système doit être appris pour pouvoir reconnaître la parole d'un humain. Les conditions pour l'apprentissage des modèles acoustiques utilisés étaient plutôt bonnes : un environnement avec peu de bruit ambiant, une parole prononcée claire et forte (avec de nombreux locuteurs et aussi beaucoup de variations pour un même mot) et une distance minimale au microphone. Ces enregistrements accompagnés de leur transcription servent à fabriquer les modèles du système de reconnaissance. Une idée intéressante serait d'ajouter aussi quelques types de bruits ambiants, pour aider le système de reconnaissance à ne pas confondre les bruits avec la parole.
- les microphones
  Il ne parait pas envisageable d'obliger les résidents à porter un micro-casque, il faut donc disposer des microphones dans la maison. Pour ce faire, la maison devrait être équipée de microphones de bonne qualité, qui seraient en écoute permanente, prêts à satisfaire les désirs des résidents. Leur emplacement doit être le plus éloigné possible des sources de bruits potentiels (appareils électroménagers, ...), mais rester compatible avec les habitudes de vie des résidents. Le plus grand problème pour la performance des microphones est lié à l'environnement dans lequel les sons sont acquis. Idéalement, les murs des pièces du bâtiment devraient être « plats », avec le moins de surfaces réfléchissantes pour éviter, le plus possible, la réverbération[3, 11, 20, 21].
- le bruit ambiant, la variabilité de la parole et la prise de son distante
  Tous ces aspects posent des problèmes dans la reconnaissance de la parole. Le système doit être indépendant au locuteur, il doit bien comprendre les mots prononcés par différents locuteurs (ou par le même locuteur dans différents états émotionnels). Le système devrait aussi être capable de reconnaître la parole même si l'enregistrement des microphones est perturbé par des bruits ambiants ou si le locuteur est loin de ceux-ci.
- détection parole / non parole
  Au sein d'un tel système, les microphones « écoutent » en permanence tout ce qui se passe dans la maison. Chaque bruit, chaque son ou chaque parole va donc être capté. Comme le système doit extraire des commandes à partir de sons et contrôler les périphériques dans le domicile en fonction des instructions obtenues, il sera nécessaire de bien détecter (et isoler) la parole pertinente parmi tous les sons environnants. Pour ce faire, on doit segmenter le signal audio en « parole, musique, bruit ». Il existe différentes méthodes de discrimination, dont la séparation du bruit et des parties voisées, la détection des discontinuités harmoniques et l'extraction de la composante parole [14, 23, 6].
- différencier la parole
  Il faut aussi que le système fasse la différence entre les commandes adressées à la centrale domotique et les conversations (résidents discutant entre eux, ...), il parait donc nécessaire d'utiliser un mot-clé au début de chaque commande pour demander l'attention du système de reconnaissance.

Un tel mot-clé doit être :

- long et sonore
  Un mot long et sonore au début d'une commande domotique peut améliorer la performance d'un système de reconnaissance. Il sera ainsi plus facile à reconnaître.
- peu commun
  On doit éviter de choisir un mot utilisé très fréquemment dans le langage courant.

– augmenter le plus possible la performance
  Toutes les méthodes et techniques que nous allons utiliser dans le système de reconnaissance sont très fortement paramétrables. Nous allons donc devoir essayer de nombreuses configurations, paramètres de reconnaissance et modèles de langage différents afin de trouver les conditions qui vont permettre d'obtenir des performances optimales (cf. chapitre 4).

## 1.4 Objet du stage

Le but du stage est d'évaluer les performances de reconnaissance dans le contexte domotique. Dans cette idée on a suivi les étapes :

– mise en place d'une chaîne d'acquisition pour créer une base de signaux collectés à distance (correspondants à une liste de commandes domotiques)
– adaptation des modèles avec des données enregistrées aussi à distance
  Pour cela on exploite des données de parole pour lesquelles on dispose de transcriptions ; on joue les données par haut parleur, et on enregistre le signal à distance. Ensuite, les transcriptions des signaux originaux sont synchronisées avec les mêmes signaux enregistrés à distance (cf. sous-chapitre 2.3).
– puis analyse de l'influence de divers paramétrages du systèmes de reconnaissance.

# Chapitre 2

# Reconnaissance de la parole

Les caractéristiques du signal de parole que l'on doit prendre en compte pour la reconnaissance de la parole sont :

- *la continuité* : Le discours oral est entendu comme une suite des mots liés entre eux, ce qui ne permet pas au module d'analyse du signal de trouver a priori la séparation entre les mots successifs. Pour cela, on n'envoie pas au moteur de reconnaissance les mots prononcés, mais les énoncés qui représentent le flux de la parole entre deux moments de silence. Un énoncé peut contenir un ou plusieurs mots. Le mot, à son tour, est composé de sons élémentaires, les phonèmes. Le *phonème* est l'unité linguistique distinctive minimale qui permet une description simplifiée d'une langue.
- *la variabilité* : On peut associer plusieurs prononciations à un même mot, selon les locuteurs ou les différents états d'un même locuteur.
- *l'encodage* : La reconnaissance de la parole n'est pas limitée strictement aux informations véhiculées par le signal audio. On peut donc trouver dans la parole des informations extralinguistiques, concernant le sujet du discours, la relation entre les locuteurs, leur état émotionnel, etc.

## 2.1 Analyse du signal vocal

A partir d'un signal vocal, on extrait un ensemble de paramètres pertinents afin de minimiser les temps de traitement et l'encombrement en mémoire. Le paramétrage du signal est effectué sur des trames successives de signal de courte durée (pour lesquelles on peut considérer le signal comme étant quasi stationnaire, typiquement 32 ms). Les trames successives se recouvrent et le décalage entre deux trames successives est typiquement de 10 ms [13].

Le signal vocal est représenté par une séquence d'ondes dans le système fréquence – amplitude. La fréquence indique le ton, et l'amplitude le volume. Dans le cas où les sons sont transmis par téléphone, l'échantillonnage est fait à ~8 kHz, dans le cas où les sons sont captés par un microphone, l'échantillonnage est fait à ~16 kHz.

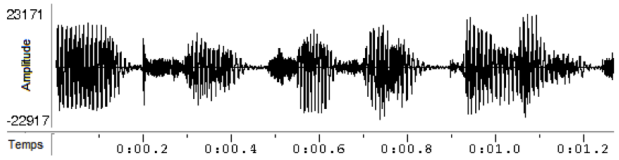



FIGURE 2.1 – Exemple de signal acoustique $s(t)$.

Plusieurs traitements sont effectués lors de l'analyse du signal acoustique :

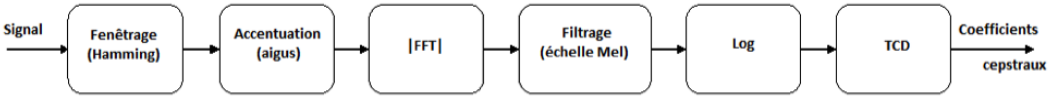



FIGURE 2.2 – Principe de l'analyse cepstrale.

– le fenêtrage de Hamming est utilisé afin de réduire les discontinuités dans le signal. Elle est définie par l'équation :

$$h(n) = \begin{cases} 0.54 - 0.46cos(2\pi\frac{n}{N-1}),\ si\ 0 \leq n \leq N-1 \\ 0,\ sinon \end{cases}$$

où N est la taille de la fenêtre. Le signal devient :

$$s_1(t) = s(t) * h(t)$$

– une pré-accentuation est appliquée pour privilégier les sons aigus :

$$s_2(t) = s_1(t) - 0.98 \cdot s_1(t-1)$$

– l'analyse de Fourier est appliquée sur chaque trame limitée par la fenêtre de Hamming $h$.

$$S_3(f) = \sum_{n=0}^{N-1} s_2(n) e^{-i2\pi fn}$$

Cela donne un spectre fréquentiel à court terme. Après avoir fait glisser la fenêtre et concaténé les spectres à court terme successifs, on obtient le spectrogramme qui montre l'évolution temps-fréquence du signal. Les zones sombres indiquent des maxima d'énergie.

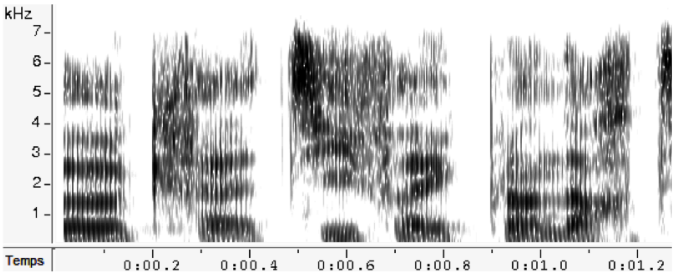



FIGURE 2.3 – Le spectrogramme d'un signal audio de parole.

– filtrage par banc de filtres

Les filtres triangulaires uniformément espacés sur l'échelle Mel (cf. figure 2.4) réduisent le nombre de bandes de fréquence, par rapport à la $FFT$ (*Transformée de Fourier*) et modélisent la non-linéarité de la perception audio humaine au niveau des fréquences. Chaque filtre fournit un coefficient qui donne l'énergie du signal dans la bande qu'il couvre.

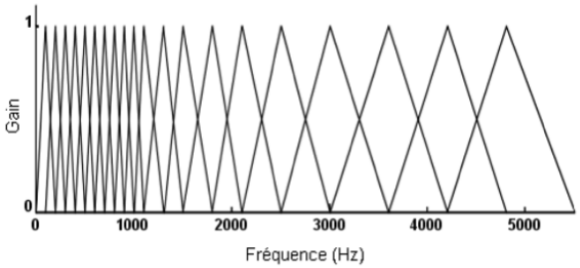



FIGURE 2.4 – Banc de filtres à échelle Mel.
On donne comme exemple le cas d'un banc de 20 filtres. Les 10 premiers ont leurs fréquences centrales distribuées linéairement. Les 10 suivants sont distribués en échelle logarithmique.

– domaine log-spectral

Le passage dans le domaine log-spectral permet de déconvoluer le signal.

– transformée en cosinus discrète

On calcule un ensemble de M coefficients (en général, M choisit entre 10 et 15) selon la transformée en cosinus discrète :

$$c_i = \sum_{j=0}^{N_f} S(j) cos(i(j - \frac{1}{2})\frac{\pi}{N_f}) \, pour \; i = 0, \, 1, \ldots, \, M$$

où Nf indique le nombre de filtres utilisés, et $S(j)$ est le logarithme de l'énergie dans le filtre $j$.

L'application de ces traitements permet d'obtenir les coefficients décorrélés MFCC (en anglais *Mel Frequency Cepstral Coefficients*). Seuls les 12 premiers coefficients sont retenus. Leur dérivées temporelles premières et secondes, liées à la vitesse et à l'accélération de la variation du spectre, permettent d'obtenir une amélioration sensible des performances. On ajout aussi le logarithme de l'énergie de la trame et ses dérivées première et seconde, pour un total de 39 paramètres.

Un type particulier d'analyse cepstrale est l'analyse MFCC Aurora, qui avait été développée pour la reconnaissance distribuée dans le cadre d'une normalisation à l'ETSI (*European Telecommunications Standards Institute*). La reconnaissance distribuée spécifie l'analyse acoustique effectuée dans un terminal (par exemple terminal mobile) et la reconnaissance effectuée sur un serveur (i.e. système centralisé). Seuls les coefficients acoustiques sont transmis dans cette approche, entre le terminal et le serveur [7]. Dans un premier temps l'analyse acoustique a été rendue robuste au bruit en y incluant une étape de débruitage et d'adaptation à la ligne [10]. Les travaux suivants (dans le cadre de l'ETSI) ont porté sur l'introduction du calcul du pitch (nécessaire pour le traitement des langues tonales, et aussi pour reconstruire un signal de parole compréhensible à partir des coefficients acoustiques) [18].

## 2.2 Modélisation pour la reconnaissance de la parole

Le processus de reconnaissance automatique de la parole doit trouver la séquence la plus vraisemblable de mots $W$ étant donnée une séquence d'observations acoustiques $O$, c'est-à-dire

$$\underset{W}{ArgMax}\, P(W|O)$$

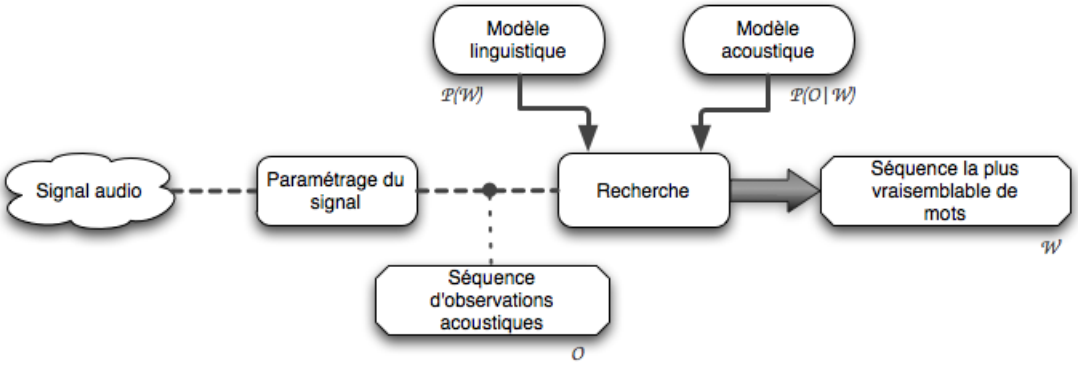



FIGURE 2.5 – Architecture d'un système de reconnaissance vocale.

On considère les éléments suivants :

– le modèle linguistique $P(W)$

On utilise un modèle linguistique appris sur un grand corpus de texte, comprenant un grand nombre de phrases $W_i$, $1 \leq i \leq l$. Chaque phrase $W$ est composée d'un ensemble de mots, $W = w_1, w_2, \ldots, w_n$. Le modèle correspond à un processus probabiliste qui établit la probabilité d'avoir dans une certaine phrase $W$ un certain mot $x$ connaissant les mots précédents.

On détermine donc la probabilité d'avoir une phrase $W$ par :

$$P(W) = P(w_1, w_2, \ldots, w_k) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \ldots P(w_k|w_1, \ldots, w_{k-1})$$

qui est simplifiée pour les langages bi-grammes avec :

$$P(w_i|w_1, \ldots, w_{i-1}) = P(w_i|w_{i-1})$$

On a donc :

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_2) \ldots P(w_k|w_{k-1})$$

– le modèle acoustique $P(O|W)$

Décrit dans le sous chapitre 2.4.

– le théorème de Bayes pour créer un modèle génératif :

$$\underset{W}{ArgMax}\, P(W|O) = \underset{W}{ArgMax} \frac{P(O|W)P(W)}{P(O)} \equiv \underset{W}{ArgMax}\, P(O|W)P(W)$$

## 2.3 Principes de la reconnaissance des formes

Le processus classique de la reconnaissance de formes est divisé en plusieurs étapes :

- le prétraitement
- l'extraction des informations pertinentes
- la classification de la forme en fonction de la meilleure ressemblance parmi une base de formes connues.

Dans le cas général, la reconnaissance se fait juste en comparant cette forme avec les formes moyennes représentatives des classes. Mais dans le cas de la reconnaissance de la parole, on doit aussi prendre en compte le fait qu'une même forme peut subir des petites transformations qui ne doivent pas changer sa classe d'appartenance. Comme énoncé précédemment, un même mot peut être prononcé de différentes manières par des locuteurs différents ou par le même locuteur dans différents états émotionnels. En conséquence, deux formes acoustiques ne peuvent pas être comparées point à point, mais seulement avec un alignement élastique [17].

La méthode d'alignement élastique utilise la programmation dynamique (en anglais $DP$ ou *Dynamic Programming*). On veut comparer deux formes acoustiques $A$ et $B$, la forme $A$ contenant $I$ vecteurs de paramètres, la forme $B$ contenant $J$ vecteurs de paramètres. Les deux formes $A$ et $B$ sont alignées par un chemin $C = [C(k)]$ dans l'espace des formes, où $C(k) = [m(k), n(k)]$ pour $1 \leq k \leq K$, $K$ étant la longueur du chemin et $m(k)$ et $n(k)$ représentant les événements mis en correspondance.

Les contraintes sur le chemin sont les suivantes :

- les fonctions $m(k)$ et $n(k)$ doivent être croissantes et respecter les conditions de continuité
- les extrémités du chemin sont établies de façon suivante : $C(1) = [1, 1]$ et $C(K) = [I, J]$.
- on considère seulement 3 voisins lors de la mise en œuvre de la DP. On indique à chaque étape soit une insertion (parcours horizontal), soit une omission (parcours vertical), soit une substitution (parcours diagonal). La relation de récurrence est alors :

$$g(i,j) = min \begin{cases} g(i,j-1) + dE(A_i, B_j) \\ g(i-1,j-1) + 2dE(A_i, B_j) \\ g(i-1,j) + dE(A_i, B_j) \end{cases} , \; 2 \leq i \leq I, \; 2 \leq j \leq J$$

où $dE(A_i, B_j)$ donne la distance entre deux vecteurs de coefficients acoustiques (par exemple MFCC) mis en correspondance, $A_i$ et $B_j$, après la formule Euclidienne :

$$dE(x,y) = \sqrt{(x-y)(x-y)^t}$$

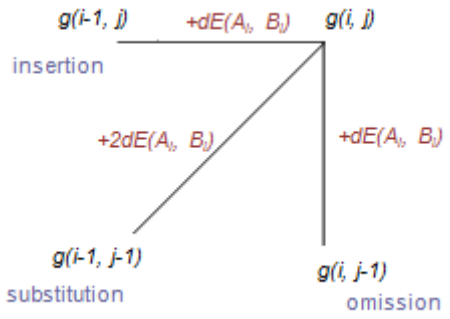



FIGURE 2.6 – Contrainte locale de recalage temporel.

Cela nous conduit à calculer la distance optimale entre $A$ et $B$ de la façon suivante :

$$D(A,B) = \frac{g(I,\ J)}{I+J}$$

– le chemin optimal peut être retrouvé par cheminement arrière, à partir de la position d'arrivée $(I,\ J)$ jusqu'au point de départ *(1, 1)*. Si les formes sont identiques, le chemin va suivre la diagonale principale de la matrice obtenue par DP.

## 2.4 Modèles de Markov

Les *modèles de Markov d'ordre 1* sont des automates probabilistes à états finis qui se basent sur l'hypothèse que « le futur ne dépend que de l'état présent ». La probabilité qu'un tel modèle soit dans l'état $i$ au temps $t$ ne dépend donc que de l'état du modèle au temps $t-1$ :

$$P(q_t|q_{t-1}, q_{t-2}, \ldots, q_1) \cong P(q_t|q_{t-1})$$

où $q_t$ est l'état du système au temps $t$ [12, 15].

Les *chaînes de Markov* sont décrites comme un simple graphe d'états auxquels on associe une fonction de transition probabiliste. A chaque pas de temps, le modèle évolue suivant la fonction de transition et passe potentiellement dans un nouvel état. L'évolution du système n'est connue qu'à travers des statistiques.

Plus formellement, un chaîne de Markov est un couple $(S,\ A)$, où :

– $S = \{1, 2, \ldots,\ N\}$ : l'ensemble fini d'états ;

– $A : SxS \rightarrow [0, 1]$ où $A_{ij} = P(q_t = j|q_{t-1} = i),\ 1 \leq i, j \leq N$ : la fonction de transition.

Dans les $modèles\ de\ Markov\ cachés$ (en anglais HMM ou *Hidden Markov Models*) on a deux processus stochastiques interdépendants. L'état du système n'est plus directement observable ; il est caché par un processus d'observation.

Le modèle HMM, $\Lambda = [\Pi,\ A,\ B]$, est défini par :

– $A$ : ensemble de transition

– $B$ : ensemble d'émission d'observations

– $\Pi$ : les probabilités initiales.

Dans la reconnaissance de la parole, les HMMs doivent fournir la réponse à la question : « Ayant un signal acoustique O, quel est la phrase la plus probable qui a été prononcée ? »

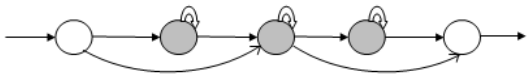


FIGURE 2.7 – Exemple de HMM à cinq états
Les états grisés sont émetteurs, les états blancs ne génèrent aucune observation.

Hypothèses simplificatrices des HMM pour son application au traitement de la parole :

- le signal de parole est produit par une suite d'états. Un modèle de Markov est associé à chaque unité de parole, leur concaténation donne les modèles des mots ou des phrases.
- le modèle est stationnaire :

$$P(q_t = j|q_{t-1} = i) \cong P(q_{t+\nu} = j|q_{t+\nu-1} = i)$$

- les observations sont indépendantes :

$$P(o_t|q_1 \dots q_t, o_1 o_2 \dots o_{t-1}) \cong P(o_t|q_1 \dots q_t)$$

- l'émission d'une observation dépend seulement de l'état courant :

$$P(o_t|q_t q_{t-1} \dots q_1) \cong P(o_t|q_t)$$

- la distribution des probabilités d'émission est approchée par un mélange de k lois gaussiennes de la forme :

$$b_j(o_t) = \sum_{i=1}^{k} \frac{c_{ji}}{\sqrt{(2\pi)^d \, |\Sigma_{ji}|}} exp(-\frac{1}{2}(o_t - \mu_{ji})' \Sigma_{ji}^{-1} (o_t - \mu_{ji}))$$

  où $\mu_{ji}$ et $\Sigma_{ji}$ sont la moyenne et la matrice de covariance de la $i^{ème}$ loi normale de la densité $b_j$, $c_{ji}$ est la pondération de la loi $i$ (avec $\sum_{i=1}^{k} c_{ji} = 1$) et $|\Sigma_{ji}|$ est le déterminant de la matrice $\Sigma_{ji}$.

La validité de ce modèle a été prouvée par les résultats.

### 2.4.1 Apprentissage

Sachant que la machine ne peut reconnaître que les formes qu'elle a déjà apprises, la fabrication des modèles de reconnaissance passe toujours par une étape d'apprentissage. La qualité de l'apprentissage est cruciale pour les performances d'un système. L'apprentissage consiste à maximiser la probabilité a posteriori qu'un modèle appris engendre bien les formes de la classe correspondante.

Le critère le plus utilisé pour l'apprentissage est le critère de *maximum de vraisemblance* (en anglais ML ou *Maximum Likelihood*). L'idée ici est de trouver l'ensemble des paramètres $\Lambda$ satisfaisant au mieux la condition suivante :

$$\underset{\Lambda}{ArgMax} \prod_{j=1}^{J} P(O_j|M_j, \Lambda)$$

Pour ce faire on utilise l'algorithme de Baum-Welch qui améliore itérativement la solution obtenue. A partir d'un modèle initial $\Lambda_0$ et d'une observation $O$, on calcule un nouveau modèle $\Lambda'$. On recommence le processus en considérant le modèle obtenu comme modèle initial, jusqu'à la validation de la condition d'arrêt (atteinte d'un optimum ou nombre maximum d'itérations). Dans cet algorithme on utilise deux variables $\xi_t(i,j)$ et $\gamma_t(i)$ :

- $\xi_t(i,j) = P(q_t = i, q_{t+1} = j|O, \Lambda)$ : la probabilité d'être dans l'état $i$ au temps $t$ et dans l'état $j$ au temps $t+1$

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\Lambda)} = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}$$

- $\gamma_t(i) = P(q_t = i|O, \Lambda)$ : la probabilité d'être dans l'état $i$ au temps $t$ sachant la séquence d'observations $O$ définie par :

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$$

On calcule les variables $\alpha_t(i)$ et $\beta_t(j)$ de la façon suivante :

- $\alpha_t(i) = P(O_1 \dots O_t, q_t = i|\Lambda)$ : la probabilité d'émettre les $t$ premières observations et d'être à l'état $i$ à l'instant $t$ connaissant le modèle
  - l'initialisation : $\alpha_1(i) = \pi_i b_i(o_1)$
  - l'induction : $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right] b_j(o_{t+1}),\ 1 \leq i \leq N,\ t = 2, \dots, T$
  - terminaison : $P(O|\Lambda) = \sum_{i=1}^{N}\alpha_T(i)$.
- $\beta_t(j) = P(O_{t+1} \dots O_T, |q_t = j, \Lambda)$ : la probabilité d'émettre les observations de $t+1$ à $T$ sachant que l'on part de l'état $j$ à l'instant $t$ et connaissant le modèle
  - l'initialisation : $\beta_T(i) = 1$
  - l'induction : $\beta_t(i) = \sum_{j=1}^{N} a_{ij}b_j(o_{t+1})\beta_{t+1}(j),\ 1 \leq i \leq N,\ t = T-1, \dots, 1$
  - terminaison : $P(O|\Lambda) = \sum_{i=1}^{N}\pi_i b_i(o_1)\beta_1(i)$.

Les lois de mises à jour des paramètres pour le nouveau modèle $\Lambda' = \left\{\overline{\Pi}, \overline{A}, \overline{B}\right\}$ sont alors :

- $\overline{\pi}_i = \gamma_1(i)$
- $\overline{a}_{ij} = \frac{\sum_{i=1}^{T-1}\xi_t(i,j)}{\sum_{i=1}^{T-1}\gamma_t(i)}$
- moyenne de la gaussienne de l'état k (dans cas monogaussian) : $\mu_k = \frac{\sum_{i=1}^{T}\gamma_t(k)o_t}{\sum_{i=1}^{T}\gamma_t(k)}$
- la covariance de la loi normale de l'état k (dans cas monogaussian) : $\Sigma_k = \frac{\sum_{i=1}^{T}\gamma_t(k)(o_t-\mu_t)(o_t-\mu_t)'}{\sum_{i=1}^{T}\gamma_t(k)}$

### 2.4.2 Reconnaissance

Dans ce sous chapitre on présente les façons de décoder la parole, et aussi l'évaluation des performances.

**1. Décodage de la parole**

Le processus de reconnaissance peut être appliqué sur des mots isolés ou de la parole continué.

*a) mots isolés* Après avoir appris les modèles des mots du vocabulaire, le processus de reconnaissance doit trouver la meilleure séquence d'états qui pourrait donner la suite d'observations correspondant à la prononciation d'un certain mot. On a :

- le vocabulaire de N mots
- les modèles des mots du vocabulaire : $M_1, M_2, \ldots, M_N$
- la séquence d'observations $O = (o_1, o_2, \ldots o_T)$.

La solution optimale de ce problème est fournie par l'algorithme de $Viterbi$ qui peut être implémenté par une matrice $TxN$ contenant les valeurs $\delta$. La vraisemblance du meilleur chemin qui finit à l'état $i$ au temps $t$ est calculée de la façon suivante :

$$\delta_t(i) = \max_{q_1,q_2,\ldots,q_{t-1}} P\left(q_1, q_2, \ldots, q_t = i, o_1, o_2, \ldots, o_t|\Lambda\right)$$

On peut effectuer ce calcul par récurrence :

- initialisation : $\delta_0(i) = \pi_i$
- hypothèse de récurrence : $\delta_t(j) = \max_i \left(\delta_{t-1}(i) \cdot a_{ij}\right) \cdot b_j\left(o_t\right)$
- terminaison : $P = \max_i \left(\delta_T\left(i\right)\right)$

Pour obtenir la séquence d'états qui a donné le maximum de vraisemblance, on doit faire le cheminement arrière en choisissant à chaque pas, l'état qui a donné le maximum courant. Cet algorithme est appliqué pour chaque modèle de mot du vocabulaire. Enfin, on choisit le mot qui donne la plus grande vraisemblance $\delta_T$.

*b) parole continue*

Pour la reconnaissance de parole continue on doit prendre en compte le fait que l'on ne connaît ni le nombre de mots que contient la phrase prononcée, ni les « frontières » de chaque mot. On peut tout de même utiliser l'algorithme précédent en y apportant une modification : après avoir atteint le dernier état d'un des mots du vocabulaire, on peut passer vers le premier état d'un autre mot de vocabulaire.

L'algorithme modifié est donc :

- initialisation : $\delta_0^k(i) = \pi_i^k$, $k$ désignant l'un des mots du vocabulaire
- hypothèse de récurrence :

  - état non initial du mot $k$ :

$$\delta_t^k(j) = \max_i \left(\delta_{t-1}^k\left(i\right) \cdot a_{ij}^k\right) \cdot b_j^k\left(o_t\right)$$

  - état initial du mot $k$ :

$$\delta_t^k(i) = max\left(\delta_{t-1}^k\left(i\right) \cdot a_{ii}^k, \max_l \left(\delta_{t-1}^l\left(etatfinal(l)\right) * P(w_k|w_l)\right)\right) \cdot b_i^k\left(o_t\right)$$

- terminaison : $P = \max_k \left(\delta_T^k\left(i\right)\right)$ qui donne le dernier état qui maximise la séquence.

Afin de retrouver la séquence de modèles, on doit mémoriser dans une structure supplémentaire, l'état qui a donné le maximum de vraisemblance à chaque pas.

**2. Évaluation des performances**

L'évaluation de la reconnaissance de la parole est donnée par le critère $WER$ (Word Error Rate) qui mesure le rapport du nombre d'erreurs de reconnaissance sur le nombre total de mots. On a trois types d'erreurs de reconnaissance : substitution (S), omission (D) et insertion (I).

$$WER = \frac{S + D + I}{N}$$

Si on prend par exemple :

- la phrase prononcée :
  *Référence* : *C' EST* le passage du groenland où on va d' abord **** ravitailler sur la côte *EST*
- et la phrase reconnue par le système :
  *Hypothèse* : ** *** le passage du groenland où on va d' abord *VOIR* ravitailler sur la côte *OUEST*
- on trouve comme erreurs : deux *omissions* (C' → **, EST → ***), une *insertion* (**** → VOIR) et une *substitution* ( EST → OUEST).

### 2.4.3 Adaptation de modèles HMM

Pour améliorer les performances de la reconnaissance de la parole, on doit adapter les modèles HMM à l'environnement d'utilisation. En utilisant une petite quantité du discours d'un nouveau locuteur (données d'adaptation), les modèles indépendants du locuteur peuvent être adaptés au le nouveau locuteur [2, 8, 9, 22].
De manière générale, les techniques d'adaptation sont de deux types :

- normalisation du signal : le flux de parole d'entrée est normalisé
- adaptation du modèle : les paramètres du modèle sont mis à jour pour améliorer la modélisation d'un nouveau locuteur ou d'un nouvel environnement.

On s'intéresse ici à la technique d'adaptation du modèle. On considère deux méthodes, l'une par maximum a posteriori (MAP) et l'autre par régression linéaire (MLLR).

**1. MAP (Maximum A Posteriori)**

Les paramètres du modèle HMM sont choisis en maximisant la probabilité a posteriori

$$\hat{\lambda} = \underset{\lambda}{argmax} P\left(O|\Lambda\right) P_0\left(\Lambda\right)$$

où $P_0\left(\Lambda\right)$ est la distribution a priori des paramètres. Grâce à la distribution a priori des paramètres du modèle, on peut obtenir des paramètres robustes même si on utilise une petite quantité de données d'adaptation.

Si la moyenne a priori est $\mu_0$, alors l'estimation MAP pour la moyenne de la gaussienne adaptée est :

$$\hat{\mu} = \frac{\tau\mu_0 + \Sigma_{t=1}^{T}\gamma_t o_t}{\tau + \Sigma_{t=1}^{T}\gamma_t}$$

où :

- $\tau$ est un méta paramètre qui contrôle la balance entre l'estimation au maximum de vraisemblance (ML) de la moyenne et sa valeur antérieure. Généralement, il est choisit dans l'intervalle [2, 20]
- $o_t$ est le vecteur au temps $t$
- $\gamma_t$ est la probabilité d'utilisation de la gaussienne à l'instant $t$.

Le désavantage du MAP réside dans le fait qu'il s'agit d'une méthode locale : elle met à jour seulement les paramètres des gaussiennes associées à des états observés.

**2. MLLR (Maximum Likelihood Linear Regression)**

Cette méthode est basée sur la supposition que les transformations pour un locuteur (ou un environnement) spécifique sont systématiques à un ensemble de gaussiennes.
Les moyennes des gaussiennes sont mises à jour d'après la transformation :

$$\hat{\mu} = A\mu + b$$

où :

- $d$ la dimension de l'observation $O$
- $A$ est une matrice $d^2$
- $b$ est un vecteur de dimension $d$

On peut aussi écrire l'équation comme $\hat{\mu} = W\xi$, où $\xi = \begin{bmatrix} 1 & \mu_1 & \cdots & \mu_d \end{bmatrix}$ est le vecteur de la moyenne étendu.

On calcule la matrice de transformation linéaire $W$ en maximisant la vraisemblance des données d'adaptation. Une itération de l'algorithme $EM$ (*Expectation-Maximisation*) suffit à estimer les valeurs de la matrice. Toutes les gaussiennes seront ensuite mises à jour en utilisant cette transformation.

Quand on dispose de peu de données d'adaptation, on estime une unique matrice $W$. Par contre, si le corpus est beaucoup plus grand, les données peuvent être divisées en plusieurs classes. Dans ce cas, l'adaptation ne se fait pas en utilisant une seule fonction, mais un nombre de fonctions égal au nombre de classes. Le nombre optimal de classes dépend de la quantité des données d'adaptation disponibles.

Les paramètres sont alors estimés indépendant pour chaque classe :

- $\hat{\mu}_i = A_1\mu + b_1,\ i\,\epsilon\, Class_1$
- $\hat{\mu}_i = A_2\mu + b_2,\ i\,\epsilon\, Class_2$
- $\ldots$
- $\hat{\mu}_i = A_k\mu + b_k,\ i\,\epsilon\, Class_k$.

# Chapitre 3

# Évaluations préliminaires sur ESTER2

La problématique du stage résidant dans l'étude de l'impact d'une prise de son distante sur les performances de reconnaissance, on a commencé par effectuer des tests sur :

- un signal original : le corpus $ESTER2$ (Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques [4]) qui a mis à disposition un ensemble d'apprentissage et un autre de développement, pour l'estimation des résultats. Le corpus de développement de 20 fichiers (d'une durée d'environ 7 heures) contient des bulletins d'information :
  - Africa 1 - radio africaine qui présente l'actualité en Afrique
  - RFI (Radio France International) - radio française qui présente l'actualité internationale
  - INTER (France INTER) - radio française qui présente toute l'actualité
  - TVME (le nouveau nom de la Radio Télévision Marocaine, RTM) - radio marocaine.

  La durée de chaque émission varie entre 15 minutes et 60 minutes (cf. tableau A.1 en annexe A). Le signal contient des zones de parole (monologues ou dialogues par téléphone, plus ou moins bruités), de musique ou de la parole sur un fond musical. Au total il y a 129 locuteurs différents identifiés.
- un signal enregistré : pour le corpus de 20 fichiers, le signal original a été enregistré en situation réelle (en mettant une attente de ~2 secondes à la fin de l'enregistrement comme mesure de sécurité) à distance avec un microphone Sony située à une distance de 1 mètre (cf. tableau A.2 en annexe A). Pour ces données enregistrées, il est nécessaire de synchroniser les transcriptions de référence par rapport au signal enregistré, pour l'adaptation des modèles acoustiques et pour les évaluations de performance.

Dans notre analyse on a suivi les étapes suivantes :

- synchronisation des transcriptions de référence du signal enregistré : nécessaires pour l'évaluation des performances de reconnaissance
- évaluation des performances : analyse de l'influence de divers paramétrages du systèmes de reconnaissance.

## 3.1 Transcriptions de référence du signal enregistré

Le corpus ESTER2 a été manuellement transcrit, les transcriptions sont stockées dans des fichiers $.stm$ qui précisent les énoncés accompagnés par l'indication des trames de signal entre lesquelles ils ont été prononcés. A partir des ces fichiers $.stm$ originaux donnés, on doit créer les transcriptions de référence du signal enregistré. On considère que les énoncés restent les mêmes et que seul les trames doivent être remplacées. On doit donc faire un alignement de signaux pour synchroniser les mêmes transcriptions avec les fichiers enregistrés.

– **Alignement élastique entre chaque signal original et son signal enregistré**

On calcule l'alignement entre deux fichiers pour bien établir les intervalles dans lesquels on doit trouver les mêmes phrases.

On rappelle ici que ces fichiers contiennent tous les deux la même information (les mêmes paroles), mais que les paroles dans les fichiers enregistrés sont couvertes par le bruit ambiant, et aussi décalées par translation temporelle à cause des mesures de sécurités.

L'alignement doit alors s'approcher de la direction de la diagonale principale (cf. sous-chapitre 2.3) ainsi que commencer et finir par des lignes droites horizontales. Les lignes droites montrent les silences au début et à la fin de l'enregistrement, la diagonale principale montre la correspondance entre les mêmes événements acoustiques.

Cependant, la programmation dynamique ne peut pas être appliquée ici juste en utilisant une matrice de dimension *I x J*, car dans notre corpus on a aussi des fichiers durant presque une heure, ce qui aurait pour effet de donner des dimensions aberrantes à stocker en mémoire (de l'ordre de 357500 x 357683 = $1.3 * 10^{11}$). En tenant compte des caractéristiques des résultats qui doivent être obtenus par la DP, on a choisi d'utiliser une autre matrice, $g'$, qui traverse juste la zone de la diagonale principale de la matrice complète, $g$. La dimension de $g'$ va être *J x 400* (on suppose la déviation maximale, par rapport à la diagonale principale, de 200 trames, ce qui donne une largeur totale de $nc = 400$ trames, l'équivalent de 4 secondes). Les cellules qui n'appartiennent pas au zones $g \cap g'$ sont considérés comme contenant des valeurs infinies.

Les coordonnées $(rx, ry)$ de la matrice $g$ vont devenir $\big(ry, rx + \frac{nc}{2} - ry\big)$dans la matrice $g'$.

Les coordonnées $(x, y)$ de la matrice $g'$ vont devenir $\big(y - \frac{nc}{2} + x, x\big)$ dans la matrice $g$.

La relation de récurrence utilisée précédemment :

$$g(i,j) = min \begin{cases} g(i,j-1) + dE(A_i, B_j) \\ g(i-1,j-1) + 2dE(A_i, B_j) \\ g(i-1,j) + dE(A_i, B_j) \end{cases} , \; 2 \leq i \leq I, \; 2 \leq j \leq J$$

va donc devenir :

$$g'(ri,rj) = min \begin{cases} g'(ri-1, rj+1) + dE(A_{rj-\frac{nc}{2}+ri}, B_{ri}) \\ g'(ri-1, rj) + 2dE(A_{rj-\frac{nc}{2}+ri}, B_{ri}) \\ g'(ri, rj-1) + dE(A_{rj-\frac{nc}{2}+ri}, B_{ri}) \end{cases} , \; 2 \leq ri \leq J, \; 2 \leq rj \leq 400$$

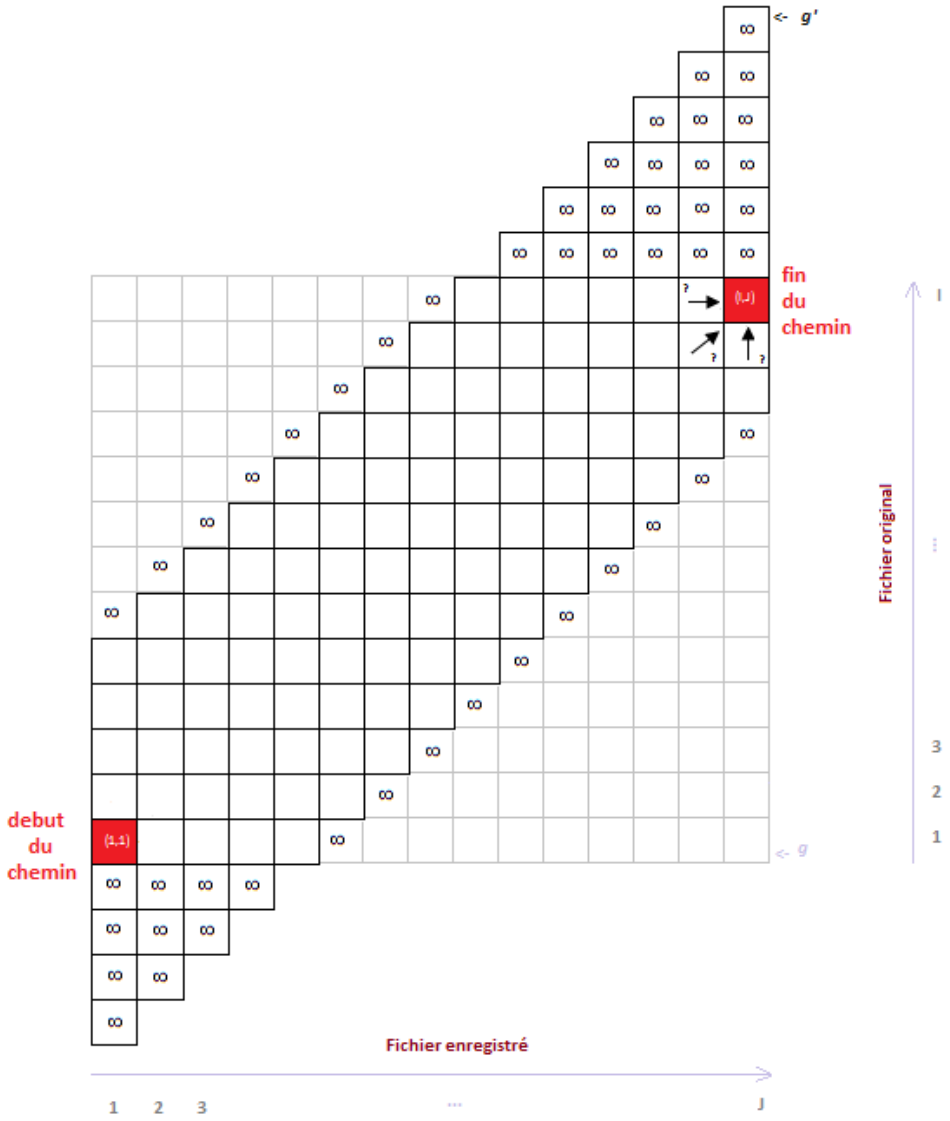



FIGURE 3.1 – Structure pour la programmation dynamique pour traites de grands fichiers.

Pour faire le cheminement arrière du point *(I, J)* jusqu'au point *(1, 1)* on a utilisé une matrice auxiliaire qui sauvegarde dans chaque cellule une valeur indiquant la direction où se trouve la valeur qui a donné le minimum courant (par exemple : la valeur 1 indique la direction horizontale, la valeur 2 – direction diagonale et la valeur 3 – direction verticale). Ainsi, on obtient le chemin qui est représenté par une liste de positions *(x, y)*, où la première position est *(I, J)*, et la dernière *(1, 1)*.

Pour visualiser et valider l'alignement temporel, on a développé un applet Java qui dessine le chemin indiqué suivant la liste des points obtenus par la DP.

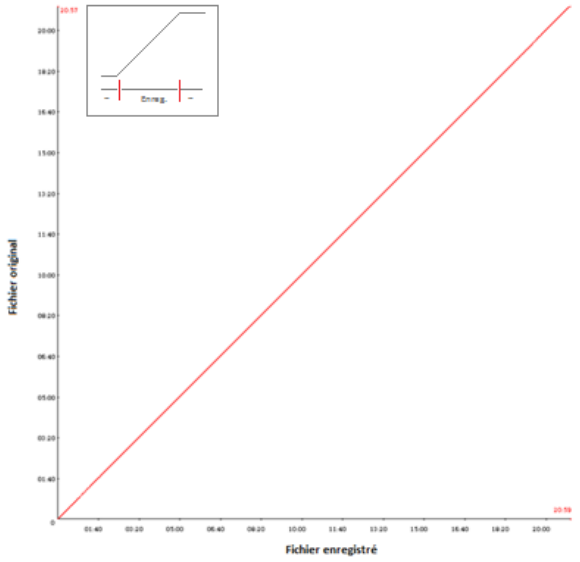



FIGURE 3.2 – Chemin du recalage entre deux formes acoustiques.
(par exemple, entre 20070710_1900_1920_inter.wav et
20070710_1900_1920_inter.sony.100cm.record.wav)

Même si le chemin semble être parfait en petite résolution, il peut cacher des « escaliers » qui rendent le processus de transposition des marqueurs temporels des transcriptions moins précis.

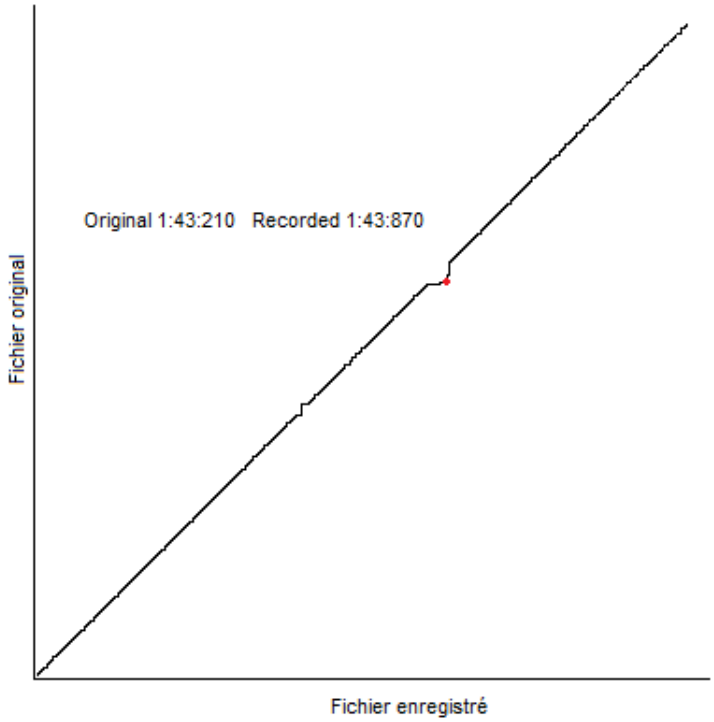



FIGURE 3.3 – Agrandissement du chemin du recalage C entre deux formes acoustiques.
(par exemple, entre 20070710_1900_1920_inter.wav et
20070710_1900_1920_inter.sony.100cm.record.wav)

L'alignement temporel entre les fichiers est très important pour le processus de reconnaissance, ce qui nécessite d'avoir une très bonne précision pour la mise en correspondance des événements acoustiques. La programmation dynamique trouve le meilleur chemin entre ces signaux audio, cependant le bruit dans le signal enregistré et les erreurs techniques ne le rendent pas idéal vis-à-vis de notre objectif.

On a fait aussi une analyse manuelle avec le programme *Wavesurfer* (un outil Open-Source permettant la visualisation et la manipulation du son) pour évaluer si le décalage entre les fichiers enregistrés et les fichiers de référence était régulier ou non. On a repéré au sein de chaque fichiers un mot ou une expression, et on a noté précisément les instants correspondants dans le fichier original et aussi dans le fichier enregistré. On a répété ce processus en choisissant un mot/expression au début, au premier, second et troisième quart, et à la fin de chaque fichiers. Les résultats montrent un décalage de l'ordre de 20 à 150 ms pour la position des mots entre fichier enregistré et fichier original. De plus, on a observé pour quelques fichiers une erreur technique constituant une perte de 60ms (6 trames) pendant l'enregistrement (cf. tableau A.3en annexe A).

Pour être sûr qu'il s'agissait bien d'une perte de signal, on a fait une analyse plus détaillée sur deux situations où on a observé l'erreur (par exemple, les fichiers $20070613 - 0730 - 0745 - africa1$ et $20070710 - 1900 - 1920 - inter$). En faisant une recherche par dichotomie, on a bien trouvé les zones où il existe une perte de signal (cf. tableau A. 4 en annexe A).

Une version visuelle des pertes du signal citées précédemment :

- dans le fichier 20070613_0730_0745_africa1

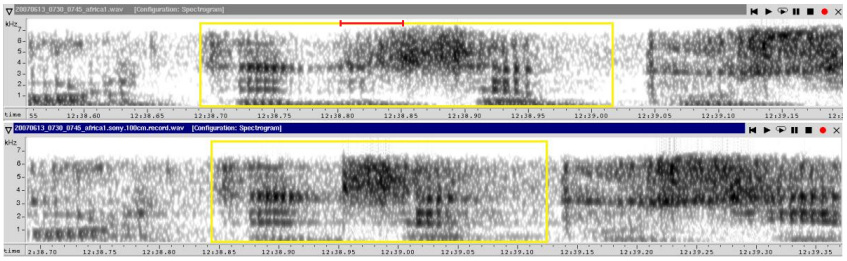


FIGURE 3.4 – Visualisation de la perte du signal dans le fichier 20070613_0730_0745_africa1.

- dans le fichier 20070710_1900_1920_inter

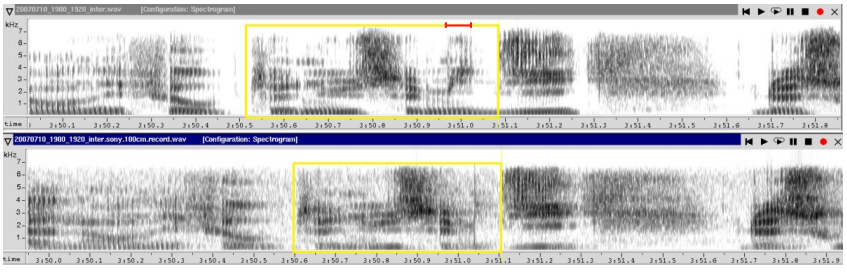


FIGURE 3.5 – Visualisation de la perte du signal dans le fichier 20070710_1900_1920_inter.

En conséquence, une autre amélioration doit être effectuée sur la détermination du chemin obtenu afin de trouver les bons intervalles pour les énonciations dans le signal enregistré. On a choisi de modifier le chemin de façon à le rapprocher de la ligne qui passe au plus près de la diagonale principale.

Pour ce faire, on a suivi les étapes suivantes :

- ignorer les extrémités du chemin qui correspondent aux moments du silence« début » et « fin » du fichier enregistré

- estimer les écarts minimum et maximum entre $x_i$et $y_i$

- garder les points qui sont les plus proches de la diagonale et ignorer les autres.

On calcule alors la régression linéaire *y=ax+b* sur les points sélectionnés $(x_i, y_i)$ en minimisant la somme des carrés des écarts des points à la droite $S = \sum_{k=1}^{K} (y_i - ax_i - b)^2$.

On doit alors résoudre le système d'équations suivant :

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

nous obtenons donc :

$$\begin{cases} a = \frac{K \cdot \sum x_i y_i - \sum y_i \cdot \sum x_i}{K \cdot \sum (x_i)^2 - (\sum x_i)^2} \\ b = \frac{\sum y_i - a \sum x_i}{K} \end{cases}$$

Avec ce réalignement (cf. tableau A.5 en annexe A) on obtient un décalage régulier dans les chemins, ceci qui élimine les « escaliers » obtenus lors de l'application de l'algorithme de programmation dynamique uniquement.

Pour valider les résultats obtenus on les compare avec quelques observations faites « à la main » dans le fichier *20070608-0730-0745-africa1.wav* :

- équation obtenue par la régression linéaire : $Rec = 1.000000 * tOri + 9.456001$
- le mot « bénéficier » au
  - temps dans le fichier enregistré 03 :31.671 (tram 21167)
  - temps dans le fichier original 03 :31.579 (tram 211572)
  - validation : 21167 ≅ 1.000000 * 21157 + 9. 456001
- le mot « Africa » au
  - temps dans le fichier original 16 :52.737 (tram 101273)
  - temps dans le fichier enregistré 16 :52.829 (tram 101282)
  - validation : 101282 ≅ 1.000000 * 101273+ 9. 456001

Le nombre de trames correspondant au temps *mm :ss.xxx* se calcule de la façon suivante :
$nt = mm * 60 * 100 + ss * 100 + \frac{xxx}{10}$.

- **Génération des fichiers .stm pour les signaux enregistrés**

On génère les fichiers de transcriptions correspondants au signaux enregistrés. On garde les énoncés originaux et on remplace les trames originaux en tenant compte de l'alignement ou du réalignement précédemment établi.

## 3.2 Évaluations de performances

L'évaluation de performance est donnée par le taux d'erreur de reconnaissance au niveau des mots WER (cf. sous-chapitre 2.4.2).

On utilise principalement la configuration classif_02 pour la transcription, qui précise les traitements à effectuer : segmentation des données en fonction des locuteurs, classement d'après leur qualité (studio ou téléphone), puis décodage avec le modèle de Markov appliqué sur chacun des segments ainsi obtenus.

### 3.2.1 Analyse des performances de la reconnaissance de parole

Elle est effectuée avec un script qui compare les résultats du décodage avec les transcriptions manuelles de référence, pour :

- les fichiers originaux accompagnés par les fichiers $.stm$ de base

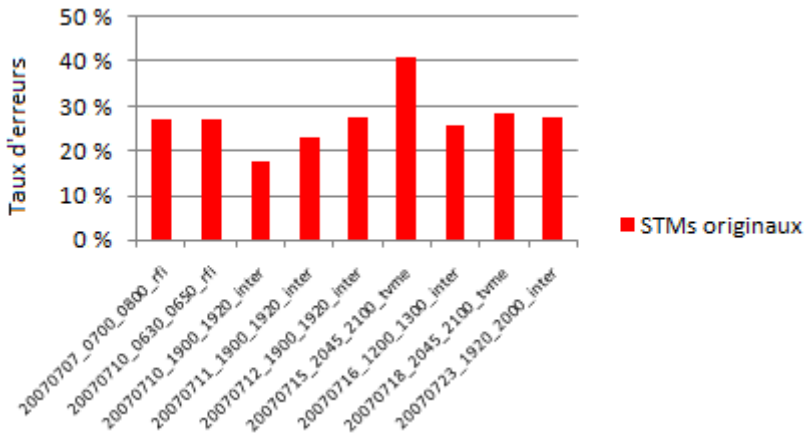



FIGURE 3.6 – Évaluation des performances de la reconnaissance sur les fichiers originaux (cf. tableau B.2 en annexe B).

- les fichiers enregistrés accompagnés par les fichiers $.stm$ originaux, les fichiers $.stm$ alignés ou les fichiers $.stm$ réalignés.

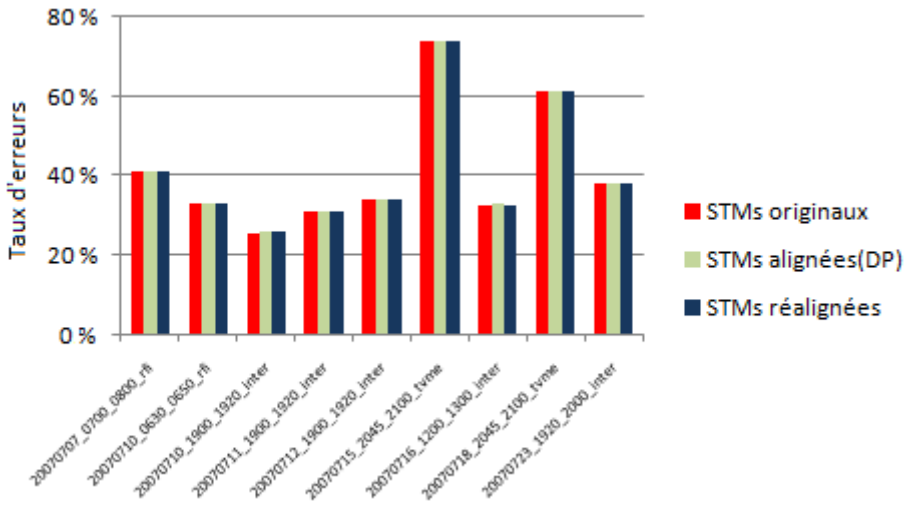



FIGURE 3.7 – Évaluation des performances de la reconnaissance sur les fichiers enregistrés (cf. tableau B.2 en annexe B).

Par contre, si on fait aussi une classification homme/femme (avec la configuration classif_01), en utilisant plusieurs jeux de modèles acoustiques adaptés à l'environnement (studio, téléphone) et au genre du locuteur (homme, femme), l'analyse des performances montre une petite amélioration dans les résultats :

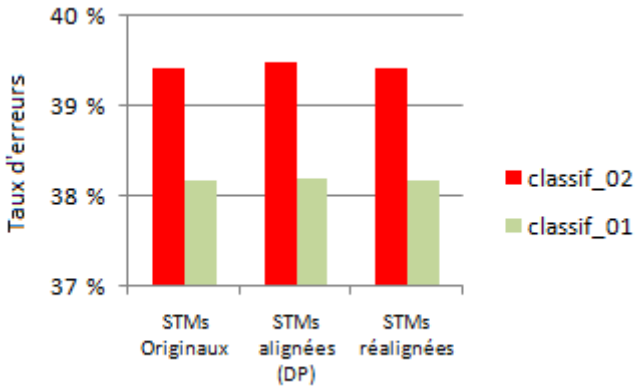



FIGURE 3.8 – Évaluation des performances de la reconnaissance avec les deux classificateurs (cf. tableaux B.1, B.2 en annexe B).

On peut observer dans les deux cas que la transcription de la parole résultant du décodage est plus proche des annotations trouvées dans les fichiers $.stm$ qui ont été réalignés par régression linéaire (classif_02 - 39.43%, classif_01 - 38.17%), que les annotations trouvées dans les fichiers $.stm$ qui ont été créés juste en utilisant le chemin obtenu par DP (classif_02 - 39.48%, classif_01 - 38.21%).

### 3.2.2 Amélioration des performances par l'adaptation des modèles HMM

On a essayé d'améliorer la performance de reconnaissance avec l'adaptation des modèles HMM. L'adaptation peut être faite de deux façons : supervisée ou non-supervisée.

**a) L'adaptation supervisée**

On rappelle que l'adaptation supervisée exploite la transcription manuelle des données d'adaptation.

On utilise comme données d'adaptation un sous-ensemble du corpus d'apprentissage d'ESTER 2 (cf. tableau A.6 en annexe A) comprenant les enregistrements radio diffusés, enregistrés sur plusieurs chaînes de radio (d'une durée d'environ 10 heures). Les données proviennent principalement de France-Inter (INTER_DGA), France-Info (INFO_DGA), Radio France International (RFI_ELDA), Radio Télévision Marocaine (RTM_ELDA), France Culture (CULTURE) et France Classique (CLASSIQUE).

Ce corpus a également été enregistré avec le même microphone Sony, situé à une distance de 1 mètre (cf. tableau A.7 en annexe A).

On fait ensuite l'adaptation des modèles HMM avec ces donnes d'adaptation en utilisant les méthodes MAP et MLLR (mentionnés dans le sous-chapitre 2.4.3).

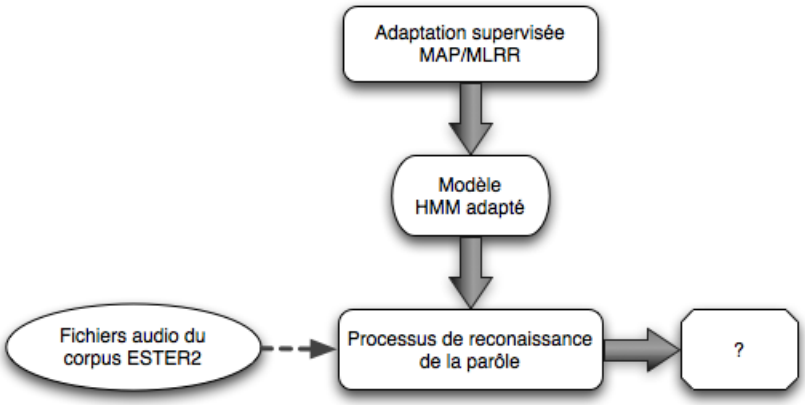



FIGURE 3.9 – Principe de l'adaptation supervisée.

Suivant cette idée, on effectue une concaténation des fichiers .stm (d'une part on fait un test avec les fichiers .stm originaux et d'autre part, avec les fichiers .stm réalignés). Les traitements nécessaires sur le fichier résultant sont :

- ajustement des annotations des bruits
- normalisation des transcriptions (addition d'espaces, passage de majuscule en minuscule)
- conversion de l'encodage du fichier en UTF8
- sélection du sous ensemble des segments du corpus en vue de l'adaptation de modèles acoustiques. La sélection est faite en fonction du lexique des prononciation disponibles. Les segments contenant des mots hors lexique sont ignorés.
- alignement sur les données de parole (pour déterminer les variantes de prononciations correspondant à chaque occurrence des mots).

On obtient les fichiers *.ctl* et *.trans*, qui précise la variante de prononciation de chaque mot de chaque segment, et qui sert pour l'adaptation supervisée des modèles :

– **MAP**
Pour l'adaptation MAP on utilise un script qui peut modifier au choix les paramètres moyennes (mean), variances (var), matrices de transition (tmat), poids de mélange (mixw) des modèles HMM. A partir du modèle HMM et du dictionnaires initiaux, et des paramètres MFCC des données d'adaptation, on va obtenir le nouveau modèle.

– **MLLR**
Pour l'adaptation MLLR on utilise :
- *bw* : le programme qui rassemble les statistiques de la moyenne et de la variance
- *mllr_solve* : le programme qui calcule la matrice de transformation
- *mllr_transform* : le programme qui transforme la moyenne.

Le modèle adapté MAP / MLLR est ensuite utilisé pour l'évaluation sur le corpus de développement d'ESTER2 enregistré à distance. On obtient les résultats suivants :

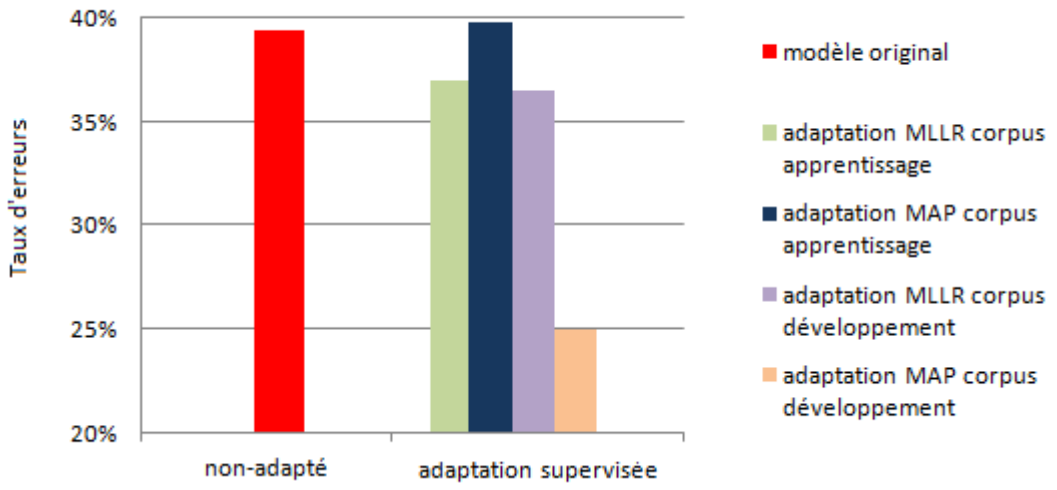



FIGURE 3.10 – Évaluation des performances de la reconnaissance après l'adaptation MLLR supervisée sur les deux corpus d'ESTER2 (cf. tableaux B.5, B.6 en annexe B).

On peut noter que le meilleur résultat obtenu après l'adaptation MAP sur le corpus d'apprentissage d'ESTER 2 (39.83%), n'améliore pas le résultat de la reconnaissance obtenu sans adaptation (39.42%). Si on effectue, par contre, une adaptation MAP sur le même corpus (c'est-à-dire sur le corpus de développement d'ESTER 2), les résultats de la reconnaissance sont meilleurs. A noter qu'une telle adaptation n'est pas réaliste en pratique. Elle permet juste d'avoir une idée des performances possibles avec une adaptation « idéale »

On peut noter aussi que l'adaptation MLLR sur le corpus d'apprentissage d'ESTER 2 (37%) améliore les résultats obtenus par rapport à l'adaptation MAP (39.83%), et aussi les résultats obtenus sans adaptation (39.42%).

Pour avoir une idée des limites de l'adaptation MLLR, on l'applique aussi sur le corpus de développement d'ESTER 2. Les résultats de la reconnaissance sont meilleurs que les résultats obtenus sans adaptation, mais plus faibles que les résultats obtenus après l'adaptation MAP.

**b) L'adaptation non-supervisée**

Dans le cas de l'adaptation non-supervisée, on fait une adaptation (MAP ou MLLR) sur les données qu'on obtient après le processus des transcription automatique, puis on relance la reconnaissance de la parole avec le nouveau modèle adapté.

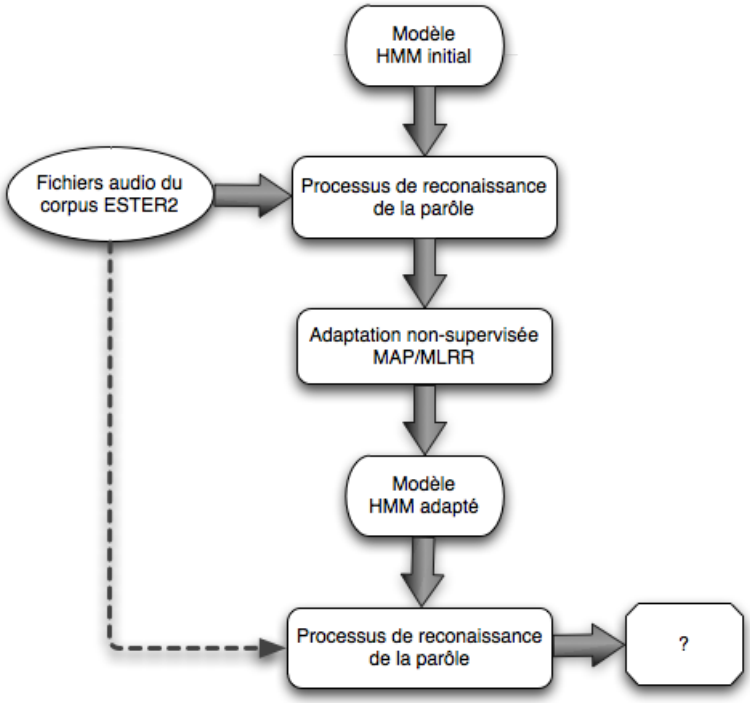



FIGURE 3.11 – Principe de l'adaptation non-supervisée.

Pour créer les fichiers $.ctl$ et $.trans$, on considère une découpe en énoncés selon les moments de silence. On fait les adaptations MAP et MLLR avec ces fichiers, ce qui donne les résultats suivants :

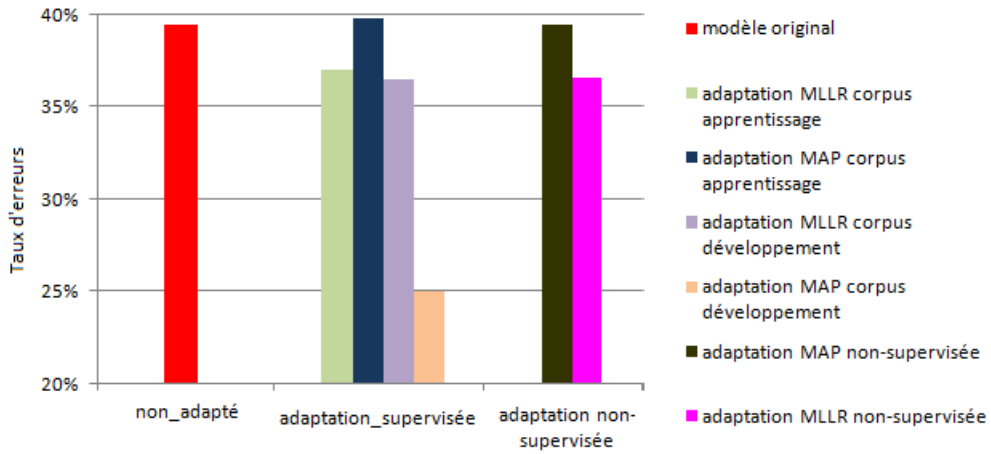



FIGURE 3.12 – Évaluation des performances de la reconnaissance sur les fichiers enregistrés (cf. tableaux B.7, B.8 en annexe B).

## 3.3 Conclusion

On rappelle ici les principaux résultats obtenus sur le corpus de développement d'ESTER2 :

- la transcription de la parole qui utilise aussi un classement de genre homme/femme donne de meilleurs performances de reconnaissance que l'utilisation d'un seul modèle générique
- la transcription de la parole résultant du décodage est plus proche des annotations trouvées dans les fichiers $.stm$ qui ont été réalignés par régression linéaire que les annotations trouvées dans les fichiers $.stm$ qui ont été créés juste en utilisant le chemin obtenu par DP
- l'adaptation des modèles HMM améliore la performance de reconnaissance. Il faut juste trouver des données d'adaptation en adéquation avec les données a reconnaître
  - l'adaptation non-supervisée par régression donne de bons résultats
  - l'adaptation MAP est sensible à la qualité de la transcription (dégradation en mode non-supervisée).

# Chapitre 4

# Évaluations sur corpus domotique

Pour évaluer la performance de la reconnaissance de la parole dans le cadre de la domotique (cf. sous-chapitre 1.3), on a mis en place une chaîne d'acquisition pour collecter des corpus de parole correspondant a une liste de commandes domotiques. Le mot clé choisit pour notre liste a été « Majordome ». On a enregistré cette liste de commandes (cf l'annexe A.2) avec deux microphones, l'un prépositionné à 40cm et l'autre à 1 mètre distance du locuteur. On a également enregistré cette liste des commandes sans le mot clé, pour vérifier que le système de reconnaissance peut bien différencier entre les commandes domotiques et le langage naturel.

Les enregistrements ont été faits dans une pièce calme et sans réverbération.

Avec le microphone prépositionné à 40cm de distance on a créé pour chaque phrase un fichier $.wav$ séparé (on appelle ces fichiers - les « fichiers segmentés »). Ce microphone était relié au logiciel d'acquisition employé (logiciel développé pour l'enregistrement de mots ou de phrases), qui affiche sur l'écran les phrases à prononcer. Le locuteur choisit, après chaque prononciation de phrase, de la répéter (si problèmes) ou de passer à la suivante.

Avec le microphone prépositionné à 1 mètre de distance on a enregistré en continu tout le discours (on appelle ces fichiers - les « fichiers en continu »).

On a collecté ces signaux prononcés par trois locuteurs, les « fichiers en continu » durant environ 1 heure (cf. tableau A.8 en annexe A).

Dans notre analyse on a suivi les étapes suivantes :

- élaboration des transcriptions de référence du signal enregistré : nécessaires pour l'évaluation de performance de reconnaissance
- évaluation de performances : analyse de l'influence de divers paramétrages du systèmes de reconnaissance.

## 4.1 Transcriptions de référence du signal enregistré

*a) Pour les « fichiers* segmentés *»*

La transcription des « fichiers segmentés » a été très facile a faire. On a pour chaque locuteur :

- un fichier $list.txt$ qui contient la liste de phrases qu'il a prononcé
- l'ensemble des « fichiers segmentés » (un fichier $.wav$ pour chaque phrase dans la liste), numérotés en concordance avec le numéro de la phrase qui a été prononcée.

Leur fichiers $.stm$ contiennent juste une ligne avec :

- l'indication du locuteur
- la trame début (0) et la trame fin (la longueur, en trames, du fichier)
- la phrase correspondant.

*b) Pour les « fichiers en continu »*

Pour faire la transcription des « fichiers en continu », un alignement doit être mis en place. Les enregistrements ont été faits simultanément avec les deux microphones. On sait alors :

- que l'ordre des phrases prononcées est le même dans les deux cas
- que la longueur d'un signal correspondant à une certaine phrase est la même dans le deux cas
- que les « fichiers en continu » peuvent contenir en plus quelques phrases mal / partiellement prononcées (erreurs d'enregistrements des « fichiers segmentés »)
- les dates de création du chaque « fichier segmenté » (sous format $hh : mm : ss$).

En exploitant ces informations, on peut chercher les meilleures correspondances (avec un alignement élastique, cf. sous-chapitre 2.3) des « fichiers segmentés » (en suivant leur ordre alphabétique) dans le « fichier en continu » correspondant, guidés par leur temps de création. Les étapes suivis sont :

- on commence avec le premier « fichier segmenté », qui a une longueur de $X$ trames. On cherche sa meilleur correspondance en glissant une « fenêtre » de $X$ trames parmi les dix premières secondes de son « fichier en continu » (les enregistrements commencent presque dans le même temps). On connaît à ce moment le temps début, $t_d$, et le temps fin, $t_f = t_d + X$, du premier « fichier segmenté » dans le cadre de son « fichier en continu ».
- on connaît aussi la date de création du premier « fichier segmenté », $t_1$, et la date de création du « fichier segmenté » suivant, $t_2$. On connaît alors le nombre de secondes écoulées entre les deux enregistrements, donné par $(t_2 - t_1)$. Il faut alors glisser une « fenêtre » de $Y$ trames (la longueur du deuxième « fichier segmenté ») parmi les prochaines $(t_2 - t_1)$ secondes du « fichier en continu », c'est-à-dire dans l'intervalle $[t_f,\ t_f + (t_2 - t_1)]$. Comme on ne connaît pas le nombre de millisecondes dans la date du création des fichiers, on peut aussi ajouter quelques trames en plus à cette intervalle comme mesure de sécurité, par exemple $[t_f + 100,\ t_f + (t_2 - t_1) + 200]$. Il est aussi réaliste de supposer une attente d'une seconde entre les prononciations.
- on répète ce processus pour tous les « fichiers segmentés » suivants.

Ces alignements permettent de créer les fichiers $.stm$ correspondants aux « fichiers en continu ». On a pour chaque commande domotique deux lignes :

- la première ligne indique les segments qui doivent être ignorés dans le processus d'analyse des résultats (de la trame $t_f$ trouvée pour le signal précédent jusqu'à trame $t_d$ trouvée pour le signal courant)
  - le format est : « locuteur 1 excluded_region $t_f$ $t_d$ <o,,unknown> ignore_time_segment_in_scoring »

– la deuxième ligne indique :
  – le locuteur
  – la trame début ($t_d$) et la trame fin ($t_f$) trouvées pour le signal courant
  – la phrase correspondant.

L'intérêt de cet alignement est de sélectionner uniquement les phrases qu'ont été bien prononcées.

On donne comme exemple le signal correspondant à la deuxième phrase prononcée par le premier locuteur (de durée 2.430 secondes), et l'adéquation dans le « fichier en continu » (dans l'intervalle [17.140, 19.570]).

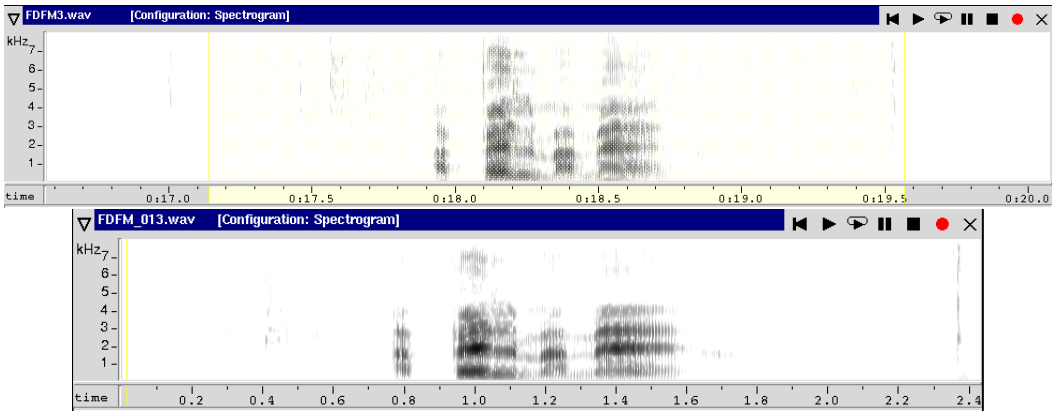



FIGURE 4.1 – Adéquation entre un « fichier segmenté » et le « fichier en continu ».

## 4.2 Évaluations de performances sur corpus domotique

Comme les données domotiques qu'on a enregistré ne sont pas suffisantes pour faire un apprentissage du système de reconnaissance, on les a utilisés uniquement comme données de test et on a conservé le modèle acoustique d'ESTER2. D'abord on utilise le modèle original, puis ensuite le modèle acoustique obtenu par adaptation avec les données d'ESTER2 enregistrées a distance.

Toutes les méthodes qu'on utilise dans le système de reconnaissance sont très fortement paramétrables. On a donc évalué de nombreuses configurations, divers paramètres de reconnaissance et modèles de langage, parmi lesquels on peut citer :

– Le vocabulaire

  Les variantes du vocabulaire sont :
  – 0.0k - on utilise juste les mots du vocabulaire domotique
  – 0.1K - on utilise les mots du vocabulaire domotique avec les 100 mots les plus fréquents du corpus ESTER2
  – 0.5K - on utilise les mots du vocabulaire domotique avec les 500 mots les plus fréquents du corpus ESTER2
  – 1K - on utilise les mots du vocabulaire domotique avec les 1000 mots les plus fréquents du corpus ESTER2

  - 5K - on utilise les mots du vocabulaire domotique avec les 5000 mots les plus fréquents du corpus ESTER2.
- Les poids modèle langage

  La formule :

$$\hat{W} \equiv \underset{W}{ArgMax}\, P(O|W)P(W)$$

  (qui indique que la solution correspond à la séquence de mots $W$ la plus vraisemblable sachant une séquence d'observations acoustiques $O$) suggère que la probabilité du modèle acoustique et la probabilité du modèle de langage peuvent être combinées à travers une simple multiplication. En pratique il est nécessaire d'effectuer une pondération. Sans cela, la participation d'un des modèles est négligeable à cause de la différence d'ordre de grandeur de leurs probabilités. En effet, les probabilités du modèle acoustique (qui sont en fait les valeurs de densités de probabilité continues multigaussiennes) sont beaucoup plus petites que celles du modèle de langage : P(A|W) « P(W). La solution la plus couramment utilisée pour atténuer ce problème consiste à ajouter un poids, noté $lw$ (pour « *linguistic weight* »), au modèle de langage [16]. On a alors :

$$\hat{W} \equiv \underset{W}{ArgMax}\, P(O|W)P(W)^{lw}$$

  Les variantes du poids modèle langage qu'on a choisit sont : 08, 07, 06, 05.
- La probabilité des fillers

  Le corpus d'ESTER2 permet de modéliser aussi des fillers (silence, respiration, bruit de bouche, bruits divers). En diminuant la probabilité des fillers on demande au système d'ajouter moins de « fillers » entre les mots.

  Les variantes du probabilités des fillers qu'on a choisit sont : 0.05, 0,01, 0.005, 0.001.
- Le modèle acoustique parametrisé du type MFCC Sphinx et MFCC Aurora

  On ajoute ici un autre type d'analyse acoustique MFCC, Aurora, plus robuste au bruit (cf. sous-chapitre 2.1), pour voir si elle apporte ou non une amélioration des résultats obtenus par rapport au paramétrage Sphinx classique. Dans notre expériences, on a conservé seuls l'énergie et les 12 premiers coefficients cepstraux de chaque trame.
- Les variantes d'adaptation sur les données d'ESTER2 jouées par haut-parleur et enregistrées à distance sont :
  - Original - modèle original (pas d'adaptation)
  - MLLR1c - adaptation type MLLR avec une seule classe
  - MLLR1cMAP - adaptation type MLLR avec une seule classe, suivie par une adaptation MAP (ré-estimation de tous les paramètres)
  - MLLR1cMAPmean - adaptation type MLLR avec une seule classe, suivie par une adaptation MAP (ré-estimation de la moyenne seulement)
  - MLLR40c - adaptation type MLLR avec 40 classes (une classe pour chaque phonème)
  - MLLR40cMAP - adaptation type MLLR avec 40 classes, suivie par une adaptation MAP (ré-estimation de tous les paramètres)
  - MLLR40cMAPmean - adaptation type MLLR avec 40 classes, suivie par une adaptation MAP (ré-estimation de la moyenne seulement)
  - MAP - adaptation MAP (ré-estimation de tous les paramètres)
  - MAPmean - adaptation MAP (ré-estimation de la moyenne seulement).

### 4.2.1 Résultats avec modèle original

**a) Impact du vocabulaire et modèle de langage associé**

La reconnaissance est appliquée uniquement sur les segments de parole indiqués dans les fichiers de transcription $.stm$. On utilise ici le modèle de langage d'ESTER2 non-adapté, le poids du modèle de langage est 08, la probabilité fillers 0.05, et les différentes variantes du vocabulaire sont : 0.0k, 0.1k, 0.5k, 1k.

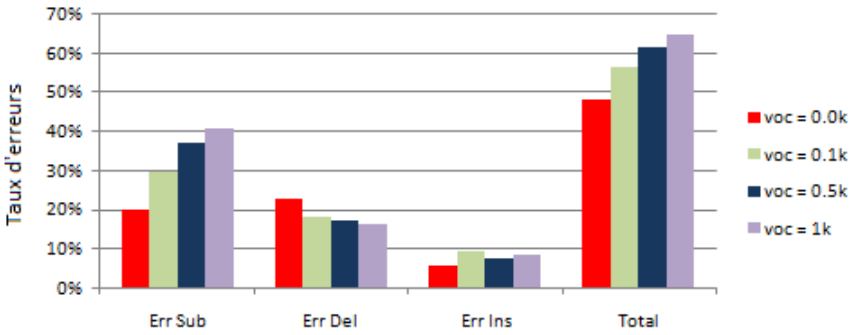



FIGURE 4.2 – Résultats avec analyse MFCC Sphinx (cf. tableau B.11 en annexe B).

Les meilleurs résultats sont obtenus avec le vocabulaire limité à celui des commandes domotiques (48.31%).

**b) Impact de quelques autres paramètres (poids du modèle de langage, probabilité des fillers)**

La reconnaissance est appliquée uniquement sur les segments de parole indiqués dans les fichiers de transcription $.stm$. On utilise ici le modèle de langage d'ESTER2, et le vocabulaire qui a donné les meilleurs résultats précédemment (0.0k).

– On varie le poids du modèle langage du 08 à 05 :

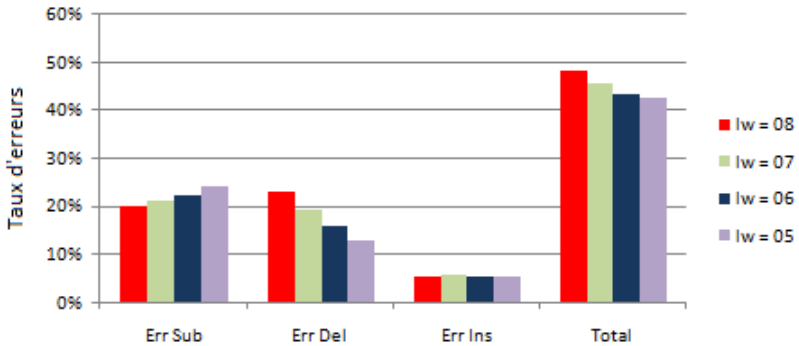



FIGURE 4.3 – Résultats avec analyse MFCC Sphinx (cf. tableau B.12 en annexe B).

En diminuant le poids du modèle langage jusqu'à 05, on améliore les résultats obtenus avant (42.56%).

– On varie les probabilités des fillers en maintenant le poids du modèle langage à 05 (valeur qui a donné les meilleurs résultats ci-dessus, 42.56%) :

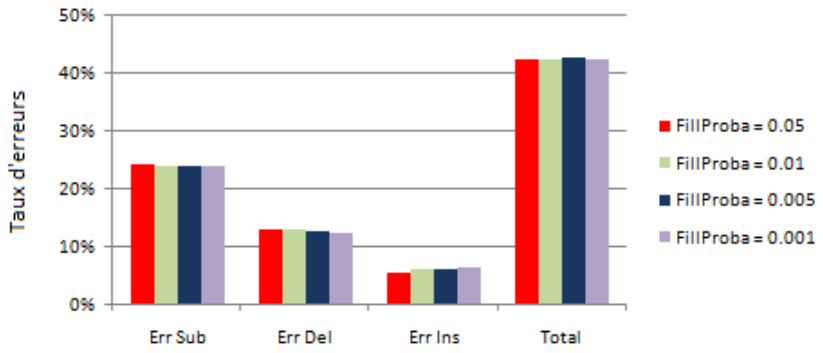



FIGURE 4.4 – Résultats avec analyse MFCC Sphinx (cf. tableau B.12 en annexe B).

La variation des probabilités des fillers n'apporte pas des gains importants, si aucun. Le meilleur résultat qu'on obtient est 42.48% (avec le probabilité des fillers égale à 0.001).

### 4.2.2 Résultats avec modèles adaptés

**a) Impact de l'adaptation des modèles acoustiques**

La reconnaissance est appliquée uniquement sur les segments de parole indiqués dans les fichiers de transcription $.stm$. On utilise ici les poids modèle langage 08, la probabilité fillers 0.05, le vocabulaire 0.0k, et les différents variantes d'adaptation du modèle acoustique sur les données d'ESTER2 enregistrées à distance.

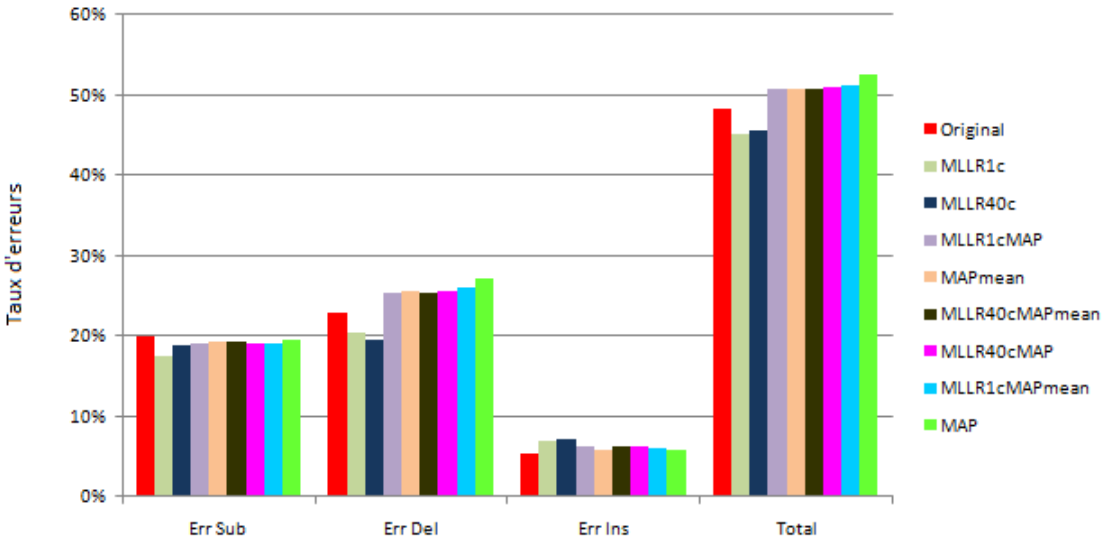



FIGURE 4.5 – Résultats avec analyse MFCC Sphinx (cf. tableau B.13 en annexe B).

Les modèles adaptés qui donnent les meilleurs résultats sont MLLR1c (45.15%) et MLLR40c (45.57%), c'est-à-dire adaptation par régression linéaire avec 1 ou 40 classes .

**b) Impact de quelques autres paramètres (poids du modèle de langage, probabilité des fillers) pour les meilleurs résultats obtenus ci-dessus**

La reconnaissance est appliquée uniquement sur les segments de parole indiqués dans les fichiers de transcription $.stm$. On utilise ici le modèle d'ESTER2 adapté type MLLR avec 40 classes et le vocabulaire limité à la domotique.

– On varie les poids modèle langage du 08 à 05 :

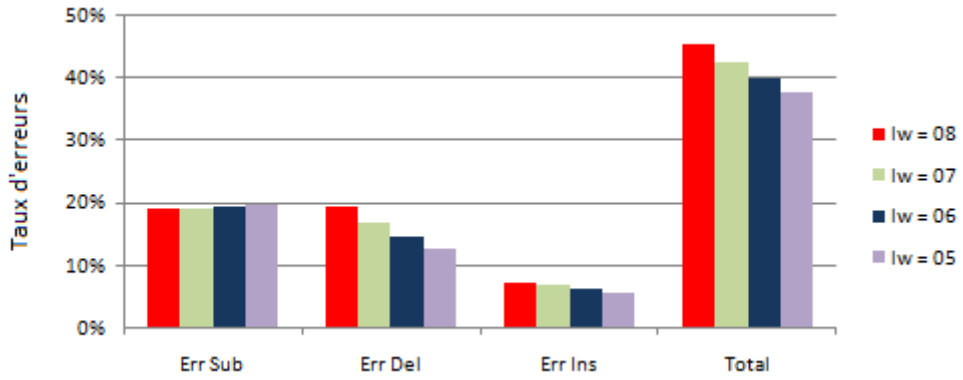



FIGURE 4.6 – Résultats avec analyse MFCC Sphinx (cf. tableau B.15 en annexe B).

En diminuant le poids du modèle de langage jusqu'à 05, on améliore les résultats obtenus avant (37.69%).

– On varie les probabilités des fillers en maintenant le poids du modèle de langage à 05 (valeur qui a donné les meilleurs résultats ci-dessus) :

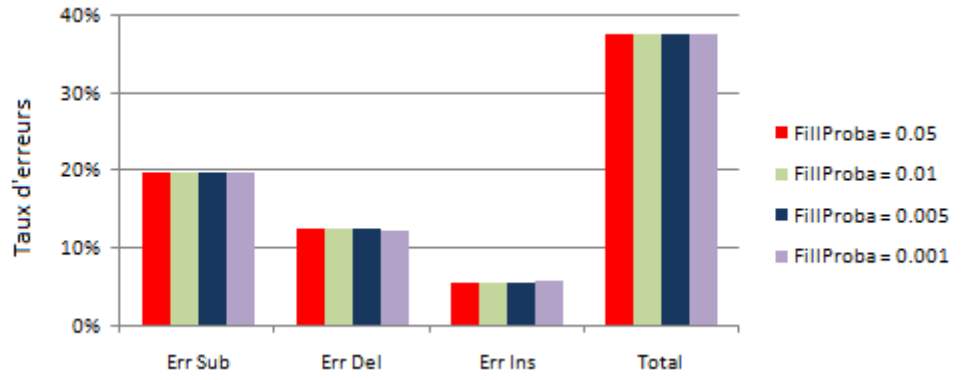



FIGURE 4.7 – Résultats avec analyse MFCC Sphinx (cf. tableau B.17 en annexe B).

La variation des probabilités des fillers n'apporte aucun gain sur les résultats.

### 4.2.3 Analyse acoustique MFCC Aurora

Dans le sous-chapitre précédant, on a présenté les meilleurs résultats que l'on peut obtenir avec l'analyse acoustique MFCC Sphinx. Nous sommes intéressés ici d'analyser la performance de reconnaissance pour le paramétrage MFCC Aurora dans les mêmes conditions.

On commence par analyser l'impact du vocabulaire et du modèle de langage associé. La reconnaissance est appliquée uniquement sur les segments de parole indiqués dans les fichiers de transcription $.stm$. On utilise ici le modèle de langage d'ESTER2 non-adapté, le poids du modèle de langage est 08, la probabilité fillers 0.05, et les différentes variantes du vocabulaire sont : 0.0k, 0.1k, 0.5k, 1k.

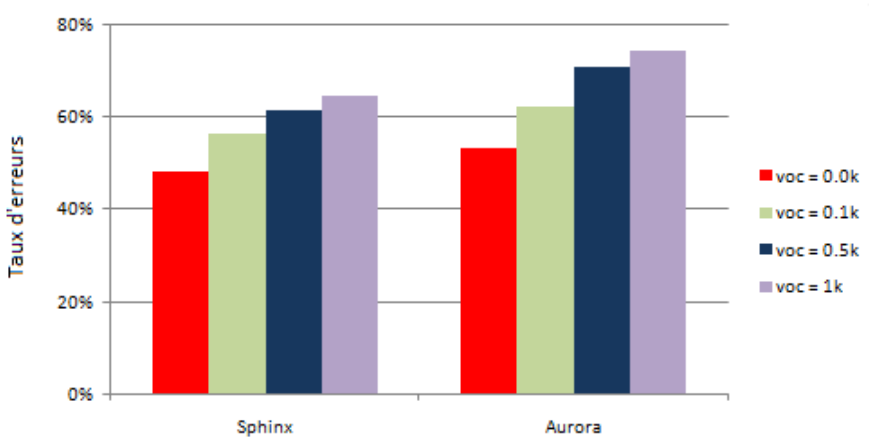



FIGURE 4.8 – Résultats avec analyse MFCC Aurora (cf. tableaux B.11 et B.12 en annexe B).

On peut observer que le paramétrage MFCC Aurora n'améliore pas les résultats obtenus avec le paramétrage Sphinx. Si on continue avec les autres types de configurations essayées dans les sous-chapitres 4.2.1 et 4.2.2, nous atteignons la même conclusion (cf. les graphiques B.1, B.2 et B.3 en annexe B).

## 4.3 Impact du mot clé

Comme mentionné avant, on a enregistré la liste des commandes domotiques avec et sans le mot clé, pour vérifier que le système de reconnaissance peut bien différencier entre les commandes domotiques et le langage naturel.

Pour évaluer l'impact du notre mot clé choisit (« majordome »), on doit faire des test en deux étapes :

- la reconnaissance est appliquée uniquement sur les segments de parole indiques dans les fichiers de transcription $.stm$ correspondants au phrases qui contiennent le mot clé
- la reconnaissance est appliquée uniquement sur les segments de parole indiques dans les fichiers de transcription $.stm$ correspondants au phrases qui ne contiennent pas le mot clé.

On utilise ici le modèle d'ESTER2 adapté type MLLR avec 40 classe, le vocabulaire limité à la domotique, le poids du modèle de langage égal à 05, et la probabilités des fillers égale à 0.005.

Les résultats obtenus pour le paramétrages MFCC Sphinx sont :

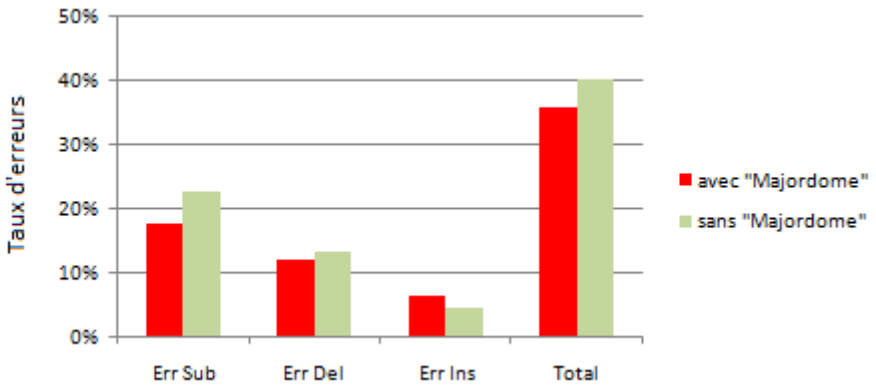



FIGURE 4.9 – Résultats avec analyse MFCC Sphinx (cf. tableaux B.19 et B.20 en annexe B).

On peut observer que l'utilisation d'un mot clé long et sonore au début de chaque phrase donne une petite amélioration des résultats. On analyse puis le taux d'erreur sur le mot « Majordome » dans les signaux audio qui contiennent le mot clé :

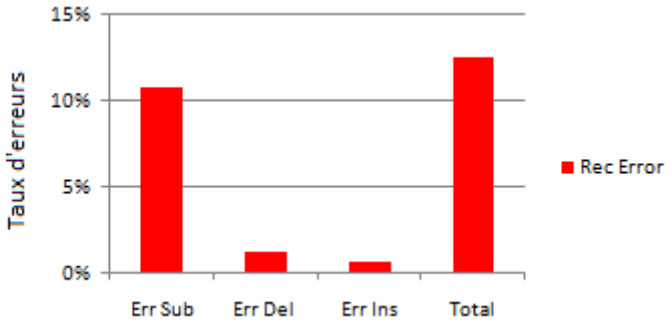



FIGURE 4.10 – Résultats avec analyse MFCC Sphinx (cf. tableau B.21 en annexe B).

On peut observer qu'on n'a pas assez d'erreurs de reconnaissance du mot « Majordome » (42 erreurs sur 336 occurrences).

## 4.4 Conclusion

On rappelle ici les principaux résultats obtenus sur le corpus domotique :

- comme les données de test ont été enregistrées dans une pièce calme et sans réverbération, le paramétrage MFCC Aurora n'apporte aucune amélioration des résultats obtenus pour le paramétrage MFCC Sphinx
- il est plus efficace d'utiliser le vocabulaire limité à la domotique, afin de ne pas augmenter le nombre d'erreurs de reconnaissance
- la diminution de la valeur du poids du modèle de langage améliore les résultats obtenus pour les deux types de paramétrage, Sphinx et Aurora. Par contre, la variation des probabilités des fillers n'apporte aucun gain important sur les résultats
- en adaptant les modèles acoustiques par régression linéaire avec 1 ou 40 classes, on améliore les résultats obtenus pour les deux types de paramétrage MFCC, Sphinx et Aurora
- un mot clé long et sonore au début de chaque commande domotique apporte des améliorations sur les résultats.

# Chapitre 5

# Évaluations sur corpus CHIME

Comme les données domotiques enregistrées dans une pièce calme et sans réverbération ne sont pas suffisantes pour évaluer la performance de reconnaissance dans des conditions moins idéales (et donc plus réelles), on a choisi d'effectuer des tests complémentaires sur le corpus disponible CHIME [1]. Ce corpus contient des données enregistrées dans un environnement domestique (salon et cuisine) réverbérant, avec plus ou moins bruits ambiants, accompagnées par leur transcription. Toutes les données audio sont stéréo, au format 16 bits WAV et échantillonnée à 16KHz. Les données d'apprentissage consistent en 17,000 énoncés réverbérés, complétées par 6 heures des bruits ambiants. Les données de développement consistent en 600 énoncés à 6 SNR (*Signal to Noise Ratio*) différents (-6dB, -3dB, 0dB, 3dB, 6dB, 9dB). Un SNR faible indique que le discours est coproduit avec un bruit de fond de haute énergie.

Ce corpus contient également les outils d'apprentissage, reconnaissance et évaluation des résultats. Avec ces outils on peut :

- estimer les paramètres du modèle HMM dépendant du locuteur (do_train_all.sh),
- effectuer la reconnaissance (do_recog_all.sh),
- évaluer la performance (score_all.sh).

L'intérêt d'avoir choisi ce corpus est de vérifier la performance du paramétrage MFCC Aurora, qui n'a apporté aucune amélioration des résultats pour les données moins bruitées.

Les résultats qui on obtient pour les parametrage MFCC HTK et Aurora sont :

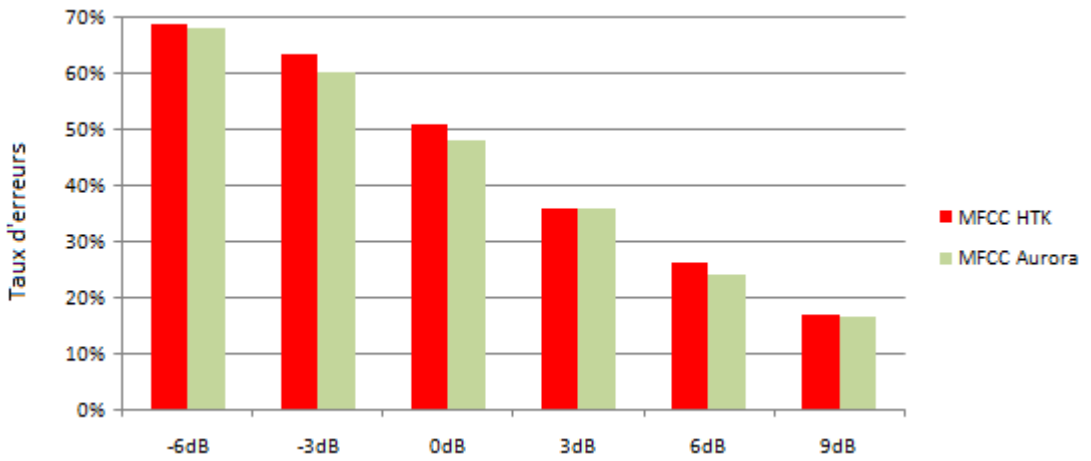



FIGURE 5.1 – Résultats avec analyse MFCC HTK&Aurora (cf. tableau B.22 en annexe B).

On peut observer que pour les données du corpus CHIME, le paramétrage MFCC Aurora apporte, comme prévu, des améliorations pour les résultats. Cela signifie que dans le cadre d'un environnement domestique, qui contient normalement des nombreux bruits ambiants, il est indiqué d'utiliser le paramétrage MFCC Aurora comme analyse acoustique du signal.

# Chapitre 6

# Conclusions

Rappelons tout d'abord l'objet du stage : on étudie les performances d'un système de reconnaissance de la parole avec prise de son distants appliqués au domaine de la domotique. Quelques questions se sont alors posées :

- comment paramétrer un signal acoustique enregistré à distance et possiblement couvert par bruit ambiant ?
- comment faire la différence entre les commandes adressées à la centrale domotique et les conversations (résidents discutant entre eux, ...) ?
- quels types de configurations, paramètres de reconnaissance et modèles de langage doivent être essayés afin de déterminer la configuration conduisant à des performances optimales du système de reconnaissance ?

Pour évaluer les performances d'un système de reconnaissance de la parole avec prise de son distante, nous avons du créer et utiliser plusieurs bases de signaux collectés à distance.

On a commencé par enregistrer à une distance de 1 mètre un sous ensemble des données de développement d'ESTER (d'une durée d'environ 7 heures). On a établi après les tests effectués sur ce corpus (cf. chapitre 3), que pour améliorer la performance de reconnaissance il faut :

- utiliser lors de la transcription une classification par environnement et genre du locuteur, et effectuer le décodage avec le modèle adéquate (spécifique à environnement et genre),
- adapter les modèles HMM au nouvel environnement (prise de son à distance ).

Ensuite, on a enregistré à une distance de 1 mètre un corpus domotique (d'une durée d'environ 1 heure). Ces données ont été enregistrées dans une pièce calme et sans réverbération, ce qui nous a mené à utiliser en plus le corpus CHIME déjà disponible pour compléter les évaluations en milieu bruité. On a établi d'après les tests effectués sur ces corpus (cf. chapitres 4 et 5), que pour améliorer la performance de reconnaissance il faut :

- utiliser le vocabulaire limité à la domotique,
- diminuer la probabilité du modèle acoustique afin de trouver la séquence de mots la plus vraisemblable, sachant une séquence d'observations acoustiques,
- adapter les modèles acoustiques par régression linéaire avec 1 ou 40 classes,
- utiliser un mot clé long et sonore au début de chaque commande domotique,
- utiliser le paramétrage MFCC Aurora comme analyse acoustique du signaux collectés dans un environnement domestique bruyant.

Pour finaliser les évaluations en contexte applicatif, il reste à aborder le problème de détection des zones parole / non parole en milieu bruyant. Il faudrait également améliorer le modèle de langage, en le fabricant à partir d'un grand nombre d'exemples de commandes domotiques.

# Annexe A

# Plus de détails sur les données utilisées

Cette annexe fournit des informations détaillées sur les données qui ont été utilisées dans notre analyse.

## A.1 Données du corpus ESTER2

1. Tableau qui précise les dimensions en octets, trames et temps des données originales de développement d'ESTER 2. Le temps $mm : ss.xxx$ correspondant au nombre de trames $NT$ se calcule après : $xxx = (NT \cdot 10)\%1000$ (nombre millisecondes), $ss = \frac{NT}{100}\%60$ (nombre secondes), $mm = \frac{NT}{100 \cdot 60}$ (nombre minutes).

| **Nom fichier** | **Dimension** ***NO*** ***(octets)*** | **Dimension** ***NT*** **(trames)** | **Temps (mm :ss.xxx)** |
|---|---|---|---|
| 20070608_0730_0745_africa1.wav | 32426154 | 101331 | 16 :53.310 |
| 20070613_0730_0745_africa1.wav | 34536848 | 107927 | 17 :59.270 |
| 20070614_0730_0745_africa1.wav | 30562892 | 95509 | 15 :55.090 |
| 20070615_0730_0745_africa1.wav | 29405982 | 91893 | 15 :18.930 |
| 20070618_0730_0745_africa1.wav | 28300898 | 88440 | 14 :44.400 |
| 20070619_0730_0745_africa1.wav | 32375164 | 101172 | 16 :51.720 |
| 20070625_0730_0745_africa1.wav | 32591666 | 101848 | 16 :58.480 |
| 20070626_0730_0745_africa1.wav | 29502112 | 92194 | 15 :21.940 |
| 20070628_0730_0745_africa1.wav | 28386998 | 88709 | 14 :47.090 |
| 20070707_0700_0800_rfi.wav | 114400046 | 357500 | 59 :35.000 |
| 20070710_0630_0650_rfi.wav | 44480046 | 139000 | 23 :10.000 |
| 20070710_1900_1920_inter.wav | 40250072 | 125781 | 20 :57.810 |
| 20070711_1900_1920_inter.wav | 40250072 | 125781 | 20 :57.810 |
| 20070712_1900_1920_inter.wav | 40260994 | 125815 | 20 :58.150 |
| 20070715_2045_2100_tvme.wav | 33920046 | 106000 | 17 :40.000 |
| 20070716_1200_1300_inter.wav | 102400046 | 320000 | 53 :20.000 |
| 20070716_2045_2100_tvme.wav | 30080046 | 94000 | 15 :40.000 |
| 20070717_2045_2100_tvme.wav | 32640046 | 102000 | 17 :00.000 |
| 20070718_2045_2100_tvme.wav | 30720046 | 96000 | 16 :00.000 |
| 20070723_1920_2000_inter.wav | 71520046 | 223500 | 37 :15.000 |
| **Somme** | **859010220** | **2684400** | **7 :27 :24.000** |

TABLE A.1 – Corpus développement ESTER 2.

2. Tableau qui précise les dimensions en octets, trames et temps des données de développement d'ESTER 2 enregistrées par le microphone Sony à une distance de 1 mètre.

| **Nom fichier** | **Dimension *NO* *(octets)*** | **Dimension *NT* (trames)** | **Temps (mm :ss.xxx)** |
|---|---|---|---|
| 20070608_0730_0745_africa1.sony.100cm.record.wav | 32473088 | 101478 | 16 :54.780 |
| 20070613_0730_0745_africa1.sony.100cm.record.wav | 34570240 | 108032 | 18 :00.320 |
| 20070614_0730_0745_africa1.sony.100cm.record.wav | 30605312 | 95641 | 15 :56.410 |
| 20070615_0730_0745_africa1.sony.100cm.record.wav | 29458432 | 92057 | 15 :20.570 |
| 20070618_0730_0745_africa1.sony.100cm.record.wav | 28344320 | 88576 | 14 :45.760 |
| 20070619_0730_0745_africa1.sony.100cm.record.wav | 32407552 | 101273 | 16 :52.730 |
| 20070625_0730_0745_africa1.sony.100cm.record.wav | 32636928 | 101990 | 16 :59.990 |
| 20070626_0730_0745_africa1.sony.100cm.record.wav | 29556736 | 92364 | 15 :23.640 |
| 20070628_0730_0745_africa1.sony.100cm.record.wav | 28442624 | 88883 | 14 :48.830 |
| 20070707_0700_0800_rfi.sony.100cm.record.wav | 114458624 | 357683 | 59 :36.830 |
| 20070710_0630_0650_rfi.sony.100cm.record.wav | 44531712 | 139161 | 23 :11.610 |
| 20070710_1900_1920_inter.sony.100cm.record.wav | 40304640 | 125952 | 20 :59.520 |
| 20070711_1900_1920_inter.sony.100cm.record.wav | 40304640 | 125952 | 20 :59.520 |
| 20070712_1900_1920_inter.sony.100cm.record.wav | 40304640 | 125952 | 20 :59.520 |
| 20070715_2045_2100_tvme.sony.100cm.record.wav | 33980416 | 106188 | 17 :41.880 |
| 20070716_1200_1300_inter.sony.100cm.record.wav | 102432768 | 320102 | 53 :21.020 |
| 20070716_2045_2100_tvme.sony.100cm.record.wav | 30113792 | 94105 | 15 :41.050 |
| 20070717_2045_2100_tvme.sony.100cm.record.wav | 32702464 | 102195 | 17 :01.950 |
| 20070718_2045_2100_tvme.sony.100cm.record.wav | 30769152 | 96153 | 16 :01.530 |
| 20070723_1920_2000_inter.sony.100cm.record.wav | 71565312 | 223641 | 37 :16.410 |
| **Somme** | **859963392** | **2687378** | **7 :27 :53.780** |

TABLE A.2 – Corpus développement ESTER 2 enregistré à distance.

3. Tableau qui détaille les décalages observés entre les fichiers originaux et les fichiers enregistrés.

| **Nom fichier** | **Mot/expression** | **Temps fichier enregistré** | **Temps fichier original** | **Décalage(ms)** |
|---|---|---|---|---|
| 20070608_0730_0745_africa1.wav | du | 00 :02.872 | 00 :02.782 | 90 |
| | bénéficier | 03 :31.671 | 03 :31.579 | 92 |
| | co | 07 :20.614 | 07 :20.522 | 92 |
| | afin | 11 :56.137 | 11 :56.042 | 95 |
| | Africa | 16 :52.829 | 16 :52.737 | 92 |
| 20070613_0730_0745_africa1.wav | tous | 00 :00.921 | 00 :00.767 | 154 |
| | per | 03 :42.044 | 03 :41.889 | 155 |
| | con | 07 :52.652 | 07 :52.498 | 154 |
| | perte | 12 :27.313 | 12 :27.158 | 155 |
| | étudier | 17 :58.614 | 17 :58.522 | 92 |
| 20070614_0730_0745_africa1.wav | de | 00 :00.694 | 00 :00.617 | 77 |
| | pour | 04 :01.759 | 04 :01.681 | 78 |
| | A | 07 :39.478 | 07 :39.461 | 17 |
| | ce | 11 :57.840 | 11 :57.822 | 18 |
| | bonjour | 15 :54.633 | 15 :54.616 | 17 |

| | | | | |
|---|---|---|---|---|
| 20070615_0730_0745_africa1.wav | bonjour | 00 :00.211 | 00 :00.115 | 96 |
| | alternative | 03 :56.596 | 03 :56.560 | 36 |
| | per | 07 :08.219 | 07 :08.182 | 37 |
| | en fait | 12 :14.824 | 12 :14.794 | 30 |
| | 7 heures | 15 :17.846 | 15 :17.809 | 37 |
| 20070618_0730_0745_africa1.wav | bonjour | 00 :00.165 | 00 :00.131 | 34 |
| | permettant | 03 :46.987 | 03 :46.952 | 35 |
| | au total | 07 :32.589 | 07 :32.558 | 31 |
| | ils sont paniques | 12 :21.394 | 12 :21.361 | 33 |
| | 7 heures | 14 :43.438 | 14 :43.406 | 32 |
| 20070619_0730_0745_africa1.wav | démocratique | 00 :00.797 | 00 :00.709 | 92 |
| | je | 03 :33.532 | 03 :33.449 | 83 |
| | pétrolier | 07 :15.591 | 07 :15.504 | 87 |
| | qui | 12 :25.744 | 12 :25.656 | 88 |
| | Africa | 16 :51.034 | 16 :50.947 | 87 |
| 20070625_0730_0745_africa1.wav | le | 00 :00.840 | 00 :00.744 | 96 |
| | a | 03 :51.253 | 03 :51.156 | 97 |
| | bien | 07 :37.840 | 07 :37.803 | 37 |
| | après ça | 11 :59.076 | 11 :59.040 | 36 |
| | Africa | 16 :57.497 | 16 :57.460 | 37 |
| 20070626_0730_0745_africa1.wav | tous | 00 :00.689 | 00 :00.629 | 60 |
| | tout | 03 :26.281 | 03 :26.221 | 60 |
| | personnes | 07 :09.477 | 07 :09.474 | 3 |
| | une | 11 :39.794 | 11 :39.793 | 1 |
| | Africa | 15 :20.877 | 15 :20.874 | 3 |
| 20070628_0730_0745_africa1.wav | en | 00 :00.926 | 00 :00.827 | 93 |
| | pos | 03 :23.363 | 03 :23.325 | 38 |
| | données | 07 :11.283 | 07 :11.245 | 38 |
| | chaque | 11 :44.958 | 11 :44.920 | 38 |
| | a | 14 :45.434 | 14 :45.394 | 40 |
| 20070707_0700_0800_rfi.wav | heure | 00 :10 :120 | 00 :10.006 | 114 |
| | de | 14 :43.747 | 14 :43.694 | 53 |
| | tour | 30 :56.891 | 30 :56.839 | 52 |
| | bien | 42 :19.069 | 42 :19.015 | 54 |
| | pour | 58 :20.378 | 58 :20.323 | 55 |
| 20070710_0630_0650_rfi.wav | trente | 00 :05.496 | 00 :05.383 | 113 |
| | et | 05 :37.147 | 05 :37.033 | 113 |
| | qui | 12 :13.208 | 12 :13.094 | 114 |
| | état | 16 :54.209 | 16 :54.095 | 114 |
| | réponde | 23 :01.465 | 23 :01.411 | 54 |

| | | | | |
|---|---|---|---|---|
| 20070710_1900_1920_inter.wav | plein | 00 :00.977 | 00 :00.896 | 81 |
| | justement | 04 :55.401 | 04 :55.380 | 21 |
| | pour | 10 :15.873 | 10 :15.853 | 20 |
| | pays | 10 :59.322 | 14 :59.301 | 21 |
| | pour | 19 :49.073 | 19 :49.053 | 20 |
| 20070711_1900_1920_inter.wav | perdu | 00 :43.208 | 00 :43.107 | 101 |
| | réponse | 04 :58.902 | 04 :58.801 | 101 |
| | une | 10 :02.596 | 10 :02.596 | 100 |
| | que | 15 :08.307 | 15 :08.206 | 101 |
| | pour | 20 :53.731 | 20 :53.630 | 101 |
| 20070712_1900_1920_inter.wav | dans | 00 :03.021 | 00 :02.924 | 97 |
| | parlamentaire | 05 :00.729 | 05 :00.631 | 98 |
| | tout | 09 :59.372 | 09 :59.335 | 37 |
| | demain | 15 :01.709 | 15 :01.671 | 38 |
| | pour | 20 :55.714 | 20 :55.676 | 38 |
| 20070715_2045_2100_tvme.wav | dimanche | 00 :17.402 | 00 :17.302 | 100 |
| | accent | 04 :00.923 | 04 :00.824 | 99 |
| | avec | 08 :01.608 | 08 :01.569 | 39 |
| | expression | 13 :01.174 | 13 :01.135 | 39 |
| | excellent | 17 :27.141 | 17 :27.101 | 40 |
| 20070716_1200_1300_inter.wav | pile | 00 :04.537 | 00 :04.429 | 108 |
| | comportement | 15 :00.772 | 15 :00.724 | 48 |
| | cameras | 29 :33.260 | 29 :33.214 | 46 |
| | instant | 43 :04.886 | 43 :04.838 | 48 |
| | termine | 53 :16.005 | 53 :15.959 | 46 |
| 20070716_2045_2100_tvme.wav | quatorze | 00 :19.983 | 00 :19.895 | 88 |
| | dix | 03 :51.126 | 03 :51.036 | 90 |
| | ancestrale | 07 :31.974 | 07 :31.885 | 89 |
| | culturel | 12 :17.698 | 12 :17.698 | 90 |
| | et | 15 :18.164 | 15 :18.135 | 29 |
| 20070717_2045_2100_tvme.wav | de | 00 :17.860 | 00 :17.779 | 81 |
| | de | 04 :13.464 | 04 :13.382 | 82 |
| | croissance | 08 :03.217 | 08 :03.136 | 81 |
| | cadre | 12 :45.639 | 12 :45.558 | 81 |
| | de | 16 :35.858 | 16 :35.777 | 81 |
| 20070718_2045_2100_tvme.wav | a | 00 :19.258 | 00 :19.148 | 110 |
| | priorité | 04 :01.096 | 04 :00.986 | 110 |
| | précède | 08 :02.565 | 08 :02.455 | 110 |
| | treize | 12 :26.028 | 12 :25.978 | 50 |
| | fidélité | 15 :36.425 | 15 :36.375 | 50 |
| 20070723_1920_2000_inter.wav | neuf | 00 :28.538 | 00 :28.507 | 31 |
| | arrive | 09 :59.402 | 09 :59.386 | 16 |
| | quelque | 20 :06.039 | 20 :06.082 | -43 |
| | paye | 29 :56.402 | 29 :56.443 | -41 |
| | très | 36 :53.144 | 36 :53.186 | -42 |

Table A.3 : Analyse manuelle avec le programme Wavesurfer.

4. Tableau qui indique en détail les décalages observés entre les fichiers originaux et les fichiers enregistrés (sur deux exemples).

| **Nom fichier** | **Mot/expression** | **Temps fichier enregistré** | **Temps fichier original** | **Décalage(ms)** |
|---|---|---|---|---|
| 20070613_0730_0745_africa1.wav | tous | 00 :00.921 | 00 :00.767 | 154 |
| | per | 03 :42.044 | 03 :41.889 | 155 |
| | con | 07 :52.652 | 07 :52.498 | 154 |
| | perte | 12 :27.313 | 12 :27.158 | 155 |
| | par tout | 12 :36.049 | 12 :35.895 | 154 |
| | de tout façon | 12 :37.627 | 12 :37.473 | 154 |
| | **hors** | **12 :38.319** | **12 :38.164** | **155** |
| | **ce** | **12 :38.953** | **12 :38.869** | **94** |
| | poursuivantes | 12.39.139 | 12 :39.045 | 94 |
| | qui | 12 :42.955 | 12 :42.861 | 94 |
| | bat | 12 :55.936 | 12 :55.841 | 95 |
| | club | 13 :41.192 | 13 :41.097 | 95 |
| | reportage | 15 :00.143 | 15 :00.049 | 94 |
| | étudier | 17 :58.614 | 17 :58.522 | 92 |
| 20070710_1900_1920_inter.wav | plein | 00 :00.977 | 00 :00.896 | 81 |
| | midi | 02 :22.139 | 02 :22.058 | 81 |
| | inter | 03 :30.899 | 03 :30.818 | 81 |
| | pour | 03 :40.862 | 03 :40.782 | 80 |
| | du | 03 :44.606 | 03 :44.526 | 80 |
| | international | 03 :47.610 | 03 :47.530 | 80 |
| | obtenue | 03 :50.608 | 03 :50.527 | 81 |
| | **midi** | **03 :50.955** | **03 :50.875** | **80** |
| | **midi** | **03 :51.104** | **03 :51.084** | **20** |
| | le | 03 :51.716 | 03 :51.695 | 21 |
| | et | 04 :10.314 | 04 :10.294 | 20 |
| | justement | 04 :55.401 | 04 :55.380 | 21 |
| | pour | 10 :15.873 | 10 :15.853 | 20 |
| | pays | 10 :59.322 | 14 :59.301 | 21 |
| | pour | 19 :49.073 | 19 :49.053 | 20 |

Table A.4 : Analyse manuelle plus détaillée avec le programme Wavesurfer.

5. Tableau qui précise les régressions linéaires du chemins obtenus par DP :

| **Nom fichier** | ***Droite*** |
|---|---|
| 20070608_0730_0745_africa1.txt | Rec = 1.000000 * tOri + 9.456001 |
| 20070613_0730_0745_africa1.txt | Rec = 1.000000 * tOri + 15.508786 |
| 20070614_0730_0745_africa1.txt | Rec = 0.999999 * tOri + 2.108969 |
| 20070615_0730_0745_africa1.txt | Rec = 1.000000 * tOri + 3.932740 |
| 20070618_0730_0745_africa1.txt | Rec = 1.000000 * tOri + 3.466521 |
| 20070619_0730_0745_africa1.txt | Rec = 1.000000 * tOri + 9.397511 |
| 20070625_0730_0745_africa1.txt | Rec = 0.999999 * tOri + 4.001478 |
| 20070626_0730_0745_africa1.txt | Rec = 1.000000 * tOri + 0.484361 |
| 20070628_0730_0745_africa1.txt | Rec = 1.000000 * tOri + 4.412619 |
| 20070707_0700_0800_rfi.txt | Rec = 1.000000 * tOri + 5.497040 |
| 20070710_0630_0650_rfi.txt | Rec = 1.000000 * tOri + 11.495674 |
| 20070710_1900_1920_inter.txt | Rec = 1.000000 * tOri + 2.463066 |
| 20070711_1900_1920_inter.txt | Rec = 1.000000 * tOri + 10.444830 |
| 20070712_1900_1920_inter.txt | Rec = 1.000000 * tOri + 4.024804 |
| 20070715_2045_2100_tvme.txt | Rec = 1.000000 * tOri + 4.095251 |
| 20070716_1200_1300_inter.txt | Rec = 1.000000 * tOri + 5.042085 |
| 20070716_2045_2100_tvme.txt | Rec = 1.000000 * tOri + 9.444146 |
| 20070717_2045_2100_tvme.txt | Rec = 1.000000 * tOri + 8.456082 |
| 20070718_2045_2100_tvme.txt | Rec = 0.999996 * tOri + 11.557738 |
| 20070723_1920_2000_inter.wav | Rec = 1.000000 * tOri - 4.196806 |

Table A.5 : Résultat de la régression linéaire.

6. Tableau qui indique les dimensions en octets, trames et temps d'un sous-ensemble des données d'apprentissage d'ESTER 2, données sélectionnées pour être enregistrées à distance.

| **Nom fichier** | **Dimension** ***NO*** ***(octets)*** | **Dimension** ***NT*** **(trames)** | **Temps (mm :ss.xxx)** |
|---|---|---|---|
| 20041006_0700_0800_CLASSIQUE.wav | 115347500 | 360460 | 60 :04.600 |
| 20041006_0800_0900_CULTURE.wav | 117802120 | 368131 | 61 :21.310 |
| 20041007_0800_0900_INTER_DGA.wav | 115387636 | 360586 | 60 :05.860 |
| 20041008_1800_1830_INFO_DGA.wav | 58333232 | 182291 | 30 :22.910 |
| 20041011_1300_1400_INTER_DGA.wav | 116196030 | 363112 | 60 :31.120 |
| 20041012_1800_1830_INFO_DGA.wav | 58175524 | 181798 | 30 :17.980 |
| 20041013_1700_1800_INFO_DGA.wav | 115377648 | 360555 | 60 :05.550 |
| 20041025_1930_2000_RFI_ELDA.wav | 57332682 | 179164 | 29 :51.640 |
| 20041026_1930_2000_RFI_ELDA.wav | 57310044 | 179093 | 29 :50.930 |
| 20041027_1230_1300_RFI_ELDA.wav | 57730406 | 180407 | 30 :04.070 |
| 20041124_1230_1300_RFI_ELDA.wav | 57995428 | 181235 | 30 :12.350 |
| 20041217_1300_1322_RTM_ELDA.wav | 42562726 | 133008 | 22 :10.080 |
| 20041218_1300_1314_RTM_ELDA.wav | 27813610 | 86917 | 14 :29.170 |
| 20041219_1300_1314_RTM_ELDA.wav | 28201964 | 88131 | 14 :41.310 |
| 20041220_1300_1314_RTM_ELDA.wav | 27440104 | 85750 | 14 :17.500 |
| 20041221_1300_1321_RTM_ELDA.wav | 40680872 | 127127 | 21 :11.270 |
| 20041222_1300_1320_RTM_ELDA.wav | 37091240 | 115910 | 19 :19.100 |
| 20041223_1300_1318_RTM_ELDA.wav | 34969644 | 109280 | 18 :12.800 |
| **Somme** | **1165748410** | **3642963** | **10 :07 :09.630** |

Table A.6 : Corpus apprentissage ESTER 2.

7. Tableau qui indique les dimensions en octets, trames et temps du sous-ensemble des données d'apprentissage d'ESTER 2 enregistrées par le microphone Sony à une distance de 2 mètres.

| **Nom fichier** | **Dimension** ***NO*** ***(octets)*** | **Dimension** ***NT*** **(trames)** | **Temps (mm :ss.xxx)** |
|---|---|---|---|
| 20041006_0700_0800_CLASSIQUE.sony.100cm.record.wav | 115408896 | 360652 | 60 :06.520 |
| 20041006_0800_0900_CULTURE.sony.100cm.record.wav | 117833728 | 368230 | 61 :22.300 |
| 20041007_0800_0900_INTER_DGA.sony.100cm.record.wav | 115441664 | 360755 | 60 :07.550 |
| 20041008_1800_1830_INFO_DGA.sony.100cm.record.wav | 58392576 | 182476 | 30 :24.760 |
| 20041011_1300_1400_INTER_DGA.sony.100cm.record.wav | 116228096 | 363212 | 60 :32.120 |
| 20041012_1800_1830_INFO_DGA.sony.100cm.record.wav | 58228736 | 181964 | 30 :19.640 |
| 20041013_1700_1800_INFO_DGA.sony.100cm.record.wav | 115408896 | 360652 | 60 :06.520 |
| 20041025_1930_2000_RFI_ELDA.sony.100cm.record.wav | 57376768 | 179302 | 29 :53.020 |
| 20041026_1930_2000_RFI_ELDA.sony.100cm.record.wav | 57344000 | 179200 | 29 :52.000 |
| 20041027_1230_1300_RFI_ELDA.sony.100cm.record.wav | 57769984 | 180531 | 30 :05.310 |
| 20041124_1230_1300_RFI_ELDA.sony.100cm.record.wav | 58032128 | 180531 | 30 :13.500 |
| 20041217_1300_1322_RTM_ELDA.sony.100cm.record.wav | 42598400 | 133120 | 22 :11.200 |
| 20041218_1300_1314_RTM_ELDA.sony.100cm.record.wav | 27852800 | 87040 | 14 :30.400 |
| 20041219_1300_1314_RTM_ELDA.sony.100cm.record.wav | 28246016 | 88268 | 14 :42.680 |
| 20041220_1300_1314_RTM_ELDA.sony.100cm.record.wav | 27492352 | 85913 | 14 :19.130 |
| 20041221_1300_1321_RTM_ELDA.sony.100cm.record.wav | 40730624 | 127283 | 21 :12.830 |
| 20041222_1300_1320_RTM_ELDA.sony.100cm.record.wav | 37126144 | 116019 | 19 :20.190 |
| 20041223_1300_1318_RTM_ELDA.sony.100cm.record.wav | 35028992 | 109465 | 19 :20.190 |
| **Somme** | **1166540800** | **3645440** | **10 :07 :34.400** |

Table A.7 : Corpus apprentissage ESTER 2 enregistré.

## A.2 Données du corpus domotique

1. Exemples de commandes domotiques

- Majordome, allume la lumière.
- Majordome, éteins la lumière.
- Majordome, baisse le niveau de la lumière.
- Majordome, augmente le niveau de la lumière.
- Majordome, ouvre les volets.
- Majordome, ferme les volets.
- Majordome, ouvre la porte d'entrée.
- Majordome, ferme la porte d'entrée.
- Majordome, baisse la température du chauffage à 19 degrés.
- Majordome, augmente la température du chauffage à 22 degrés.
- Majordome, fixe l'heure de début de la nuit à 22h.
- Majordome, fixe l'heure de fin de la nuit à 5h.
- Majordome, fixe l'heure de début de la journée à 5h.
- Majordome, fixe l'heure de fin de la journée à 22h.
- Majordome, fixe la température du chauffage pendant la nuit à 18 degrés.
- Majordome, fixe la température du chauffage pendant la journée à 21 degrés.
- Majordome, allume la climatisation.
- Majordome, éteins la climatisation.
- Majordome, fixe la température du climatiseur à 18 degrés.
- Majordome, ajoute un nouveau contact : nom Dupont, prénom Jean, téléphone 01 42 56 81 17.
- Majordome, quel est le numéro de téléphone de Jean Dupont ?
- Majordome, appelle Jean Dupont.
- Majordome, supprime le contact Jean Dupont.
- Majordome, allume la radio.
- Majordome, éteins la radio.
- Majordome, baisse le volume de la radio.
- Majordome, met le canal de radio France Inter.
- Majordome, allume la télé.
- Majordome, éteins la télé.
- Majordome, baisse le volume de la télé.
- Majordome, chaine suivante.
- Majordome, chaine précédente.
- Majordome, met la chaine TF1.
- Majordome, active le système de sécurité.
- Majordome, désactive le système de sécurité.
- Majordome, allume l'arrosage automatique.
- Majordome, éteins l'arrosage automatique.

2. Tableau précisant les dimensions en octets, trames et temps des données du corpus domotique

| **Nom fichier** | **Dimension *NO*** ***(octets)*** | **Dimension *NT*** **(trames)** | **Temps** **(mm :ss.xxx)** |
|---|---|---|---|
| FDFM | 22366044 | 69893 | 11 :38.930 |
| FDJM | 58282044 | 182131 | 30 :21.310 |
| FSRM | 40330044 | 126031 | 21 :00.310 |
| **Somme** | **120978132** | **378055** | **63 :00.550** |

Table A.8 : Corpus domotique enregistré en continu.

## A.3 Données du corpus CHIME

Tableau précisant les données du corpus CHIME

| **Données** | **Cuisine** | **Salon** |
|---|---|---|
| Enregistrements ambiantes (dev + train) | 5 :50 :08 | 5 :38 :56 |
| Enregistrements ambiantes combinés avec le corpus Grid | 16 :05 :41 | 14 :41 :29 |

Table A.9 : Corpus CHIME (la durée d'enregistrements).

# Annexe B

# Plus de détails sur les résultats

Cette annexe fournit des informations détaillées sur les résultats obtenus en utilisant differents modèles et configurations. pour le processus de la reconnaissance de la parole.

## B.1 Résultats d'ESTER 2

1. Tableau détaillant les résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_01).

| **Emission** | **Fichiers originaux** | **Fichiers enregistrés à distance** | | |
|---|---|---|---|---|
| | **STMs originaux** | **STMs originaux** | **STMs alignées (DP)** | **STMs réalignées** |
| 20070707_0700_0800_rfi | 25.58% | 40.67% | 40.67% | 40.64% |
| 20070710_0630_0650_rfi | 23.61% | 32.35% | 32.35% | 32.35% |
| 20070710_1900_1920_inter | 17.53% | 25.01% | 25.16% | 25.08% |
| 20070711_1900_1920_inter | 22.19% | 29.53% | 29.64% | 29.61% |
| 20070712_1900_1920_inter | 24.73% | 32.30% | 32.38% | 32.30% |
| 20070715_2045_2100_tvme | 39.24% | 74.05% | 74.05% | 74.05% |
| 20070716_1200_1300_inter | 23.48% | 31.22% | 31.41% | 31.20% |
| 20070718_2045_2100_tvme | 27.17% | 60.87% | 60.87% | 60.87% |
| 20070723_1920_2000_inter | 26.11% | 36.22% | 36.13% | 36.14% |
| **Moyenne** | **25.13%** | **38.18%** | **38.21%** | **38.17%** |

Table B.1. Résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_01).

2. Tableau qui précise les résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_02).

| **Emission** | **Fichiers originaux** | **Fichiers enregistrés à distance** | | |
|---|---|---|---|---|
| | **STMs originaux** | **STMs originaux** | **STMs alignées (DP)** | **STMs réalignées** |
| 20070707_0700_0800_rfi | 26.84% | 41.15% | 41.20% | 41.13% |
| 20070710_0630_0650_rfi | 26.92% | 33.38% | 33.38% | 33.38% |
| 20070710_1900_1920_inter | 17.78% | 25.91% | 26.03% | 25.98% |
| 20070711_1900_1920_inter | 23.02% | 31.30% | 31.40% | 31.40% |
| 20070712_1900_1920_inter | 27.37% | 34.14% | 34.25% | 34.14% |
| 20070715_2045_2100_tvme | 40.97% | 74.01% | 74.01% | 74.01% |
| 20070716_1200_1300_inter | 25.47% | 32.82% | 32.98% | 32.85% |
| 20070718_2045_2100_tvme | 28.18% | 61.39% | 61.44% | 61.44% |
| 20070723_1920_2000_inter | 27.67% | 38.21% | 38.17% | 38.16% |
| **Moyenne** | **26.71%** | **39.42%** | **39.48%** | **39.43%** |

Table B.2 : Résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_02).

3. Tableau détaillant les résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_02), après l'adaptation supervisée MAP du modèle HMM sur le corpus d'apprentissage d'ESTER 2 enregistré à distance.

| **Emission** | **Modèle adapté(MAP) sur le corpus d'apprentissage d'ESTER2 enregistré à distance** | | | |
|---|---|---|---|---|
| | **avec STMs originaux** | | **avec STMs réalignées** | |
| | mean,var,tmat,mixw | mean,tmat,mixw | mean,var,tmat,mixw | mean,tmat,mixw |
| 20070707_0700_0800_rfi | 43.49% | 44.02% | 43.456% | 44.20% |
| 20070710_0630_0650_rfi | 33.38% | 34.67% | 32.95% | 34.70% |
| 20070710_1900_1920_inter | 25.98% | 26.14% | 25.78% | 26.34% |
| 20070711_1900_1920_inter | 30.07% | 31.80% | 30.25% | 31.11% |
| 20070712_1900_1920_inter | 34.30% | 35.42% | 33.90% | 34.62% |
| 20070715_2045_2100_tvme | 73.83% | 72.78% | 73.94% | 72.96% |
| 20070716_1200_1300_inter | 33.19% | 34.34% | 32.98% | 33.16% |
| 20070718_2045_2100_tvme | 61.27% | 59.12% | 60.62% | 58.96% |
| 20070723_1920_2000_inter | 39.32% | 39.70% | 39.42% | 38.81% |
| **Moyenne** | **39.94%** | **40.40%** | **39.83%** | **39.99%** |

Table B.3 : Résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (après l'adaptation supervisée MAP du modèle HMM sur le corpus d'apprentissage d'ESTER 2 enregistré à distance).

4. Tableau indiquant les résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_02), après l'adaptation supervisée MAP du modèle HMM sur le corpus de développement d'ESTER 2 enregistré à distance.

| **Emission** | **Modèle adapté(MAP) sur le corpus de développement d'ESTER2 enregistré à distance** | | | |
|---|---|---|---|---|
| | **avec STMs originaux** | | **avec STMs réalignées** | |
| | mean,var,tmat,mixw | mean,tmat,mixw | mean,var,tmat,mixw | mean,tmat,mixw |
| 20070707_0700_0800_rfi | 26.95% | 28.43% | 29.49% | 27.86% |
| 20070710_0630_0650_rfi | 23,94% | 24.17% | 24.83% | 24.37% |
| 20070710_1900_1920_inter | 12.90% | 12.18% | 12.82% | 12.18% |
| 20070711_1900_1920_inter | 16.40% | 14.61% | 15.17% | 13.70% |
| 20070712_1900_1920_inter | 20.57% | 19.26% | 19.87% | 19.10% |
| 20070715_2045_2100_tvme | 69.96% | 69.92% | 69.92% | 69.47% |
| 20070716_1200_1300_inter | 15.86% | 14.77% | 14.56% | 12.67% |
| 20070718_2045_2100_tvme | 55.88% | 55.03% | 55.52% | 54.26% |
| 20070723_1920_2000_inter | 20.13% | 18.85% | 21.07% | 19.44% |
| **Moyenne** | **26.32%** | **25.32%** | **26.21%** | **24.97%** |

Table B.4 : Résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (après l'adaptation supervisée MAP du modèle HMM sur le corpus de développement d'ESTER 2 enregistré à distance).

5. Tableau qui détaille les résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_02), après l'adaptation supervisée MLLR du modèle HMM sur le corpus d'apprentissage d'ESTER 2 enregistré à distance.

| **Emission** | **Modèle adapté(MLLR) sur le corpus d'apprentissage d'ESTER2 enregistré à distance** | |
|---|---|---|
| | **avec STMs originaux** | **avec STMs réalignées** |
| 20070707_0700_0800_rfi | 39.56% | 39.59% |
| 20070710_0630_0650_rfi | 30.89% | 30.89% |
| 20070710_1900_1920_inter | 24.57% | 24.57% |
| 20070711_1900_1920_inter | 27.48% | 27.42% |
| 20070712_1900_1920_inter | 31.10% | 31.21% |
| 20070715_2045_2100_tvme | 71.20% | 71.20% |
| 20070716_1200_1300_inter | 31.46% | 31.54% |
| 20070718_2045_2100_tvme | 57.99% | 57.99% |
| 20070723_1920_2000_inter | 35.50% | 35.50% |
| **Moyenne** | **36.98%** | **37.00%** |

Table B.5 : Résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (après l'adaptation supervisée MLLR du modèle HMM sur le corpus d'apprentissage d'ESTER 2 enregistré à distance).

6. Tableau détaillant les résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_02), après l'adaptation supervisée MLLR de modèle HMM sur le corpus de développement d'ESTER 2 enregistré à distance.

| **Emission** | **Modèle adapté(MLLR) sur le corpus de développement d'ESTER2 enregistré à distance** | |
|---|---|---|
| | **avec STMs originaux** | **avec STMs réalignées** |
| 20070707_0700_0800_rfi | 38.61% | 38.66% |
| 20070710_0630_0650_rfi | 30.46% | 30.46% |
| 20070710_1900_1920_inter | 24.08% | 24.21% |
| 20070711_1900_1920_inter | 27.58% | 27.61% |
| 20070712_1900_1920_inter | 30.46% | 30.46% |
| 20070715_2045_2100_tvme | 71.12% | 71.12% |
| 20070716_1200_1300_inter | 30.48% | 30.58% |
| 20070718_2045_2100_tvme | 57.75% | 57.87% |
| 20070723_1920_2000_inter | 34.92% | 34.76% |
| **Moyenne** | **36.45%** | **36.46%** |

Table B.6 : Résultats de la reconnaissance de parole sur corpus de développement d'Ester 2 enregistré à distance (après l'adaptation supervisée MLLR du modèle HMM sur le corpus de développement d'ESTER 2 enregistré à distance).

7. Tableau qui détaille les résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_02), après l'adaptation non-supervisée MAP du modèle HMM sur le corpus de développement d'ESTER 2 enregistré à distance).

| **Emission** | **Modèle adapté(MAP) sur le corpus de développement d'ESTER2 enregistré à distance** | |
|---|---|---|
| | **avec STMs originaux** | **avec STMs réalignées** |
| 20070707_0700_0800_rfi | 41.53% | 41.15% |
| 20070710_0630_0650_rfi | 33.15% | 33.15% |
| 20070710_1900_1920_inter | 25.93% | 25.98% |
| 20070711_1900_1920_inter | 31.08% | 31.13% |
| 20070712_1900_1920_inter | 34.20% | 34.20% |
| 20070715_2045_2100_tvme | 74.62% | 74.62% |
| 20070716_1200_1300_inter | 33.19% | 33.19% |
| 20070718_2045_2100_tvme | 61.31% | 61.31% |
| 20070723_1920_2000_inter | 38.02% | 37.94% |
| **Moyenne** | **39.48%** | **39.48%** |

Table B.7 : Résultats de la reconnaissance de parole sur corpus de développement d'Ester 2 enregistré à distance (après l'adaptation non-supervisée MAP du modèle HMM sur le corpus de développement d'ESTER 2 enregistré à distance).

8. Tableau qui précise les résultats de la reconnaissance de parole sur le corpus de développement d'ESTER 2 enregistré à distance (avec la configuration classif_02), après l'adaptation non-supervisée MLLR du modèle HMM sur le corpus de développement d'ESTER 2 enregistré à distance.

| **Emission** | **Modèle adapté(MLLR) sur le corpus de développement d'ESTER2 enregistré à distance** | |
|---|---|---|
| | **avec STMs originaux** | **avec STMs réalignées** |
| 20070707_0700_0800_rfi | 38.77% | 38.74% |
| 20070710_0630_0650_rfi | 30.50% | 30.43% |
| 20070710_1900_1920_inter | 24.31% | 24.39% |
| 20070711_1900_1920_inter | 27.74% | 27.72% |
| 20070712_1900_1920_inter | 30.25% | 30.27% |
| 20070715_2045_2100_tvme | 71.50% | 71.50% |
| 20070716_1200_1300_inter | 30.45% | 30.45% |
| 20070718_2045_2100_tvme | 58.07% | 58.07% |
| 20070723_1920_2000_inter | 35.02% | 34.96% |
| **Moyenne** | **36.57%** | **36.55%** |

Table B.8 : Résultats de la reconnaissance de parole sur corpus de développement d'ESTER 2 enregistré à distance (après l'adaptation non-supervisée MLLR de modèle HMM sur le corpus de développement d'ESTER 2 enregistré à distance).

## B.2 Résultats sur corpus domotique

1. Tableau détaillant les résultats de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre.

| **Nom Fichier** | Le modèle HMM d'ESTER2 (en ignorant les silences) | Le modèle HMM d'ESTER2 (+vocabulaire domotique) |
|---|---|---|
| FDFM | 45.23% | 52.92% |
| FDJM | 82.09% | 71.01% |
| FSRM | 84.69% | 77.37% |
| **Moyenne** | **77.66%** | **70.99%** |

TABLE B.9 – Résultats de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre.

2. Tableau détaillant les résultats de la reconnaissance de parole sur le corpus domotique :
- sur les fichiers correspondants pour chaque phrase, enregistrés à une distance de ~40cm.
- sur les fichiers correspondants pour chaque phrase, extraits du fichiers enregistrés en continu.

| **Nom Fichier** | Enregistrés à 40cm distance | Extraits du fichiers enregistrés en continu à 1 mètre distance |
|---|---|---|
| FDFM | 47.81% | 49.40% |
| FDJM | 54.60% | 66.33% |
| FSRM | 65.82% | 76.32% |
| **Moyenne** | **58.39%** | **64.01%** |

TABLE B.10 – Résultats de la reconnaissance de parole sur le corpus domotique.

3. Tableaux qui détaillent l'impact du vocabulaire et du modèle de langage associé sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre

**MFCC Sphinx - Variation vocabulaire**

| Nom Fichier | Configuration : modèle original, lw = 08, FillProb = 0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 12.20% | 6.76% | 4.64% | **23.61%** | 17.24% | 7.03% | 8.62% | **32.89%** | 19.76% | 8.09% | 5.31% | **33.16%** | 20.03% | 9.15% | 4.77% | **33.95%** |
| FDJM | 20.62% | 31.60% | 5.80% | **58.02%** | 31.65% | 24.50% | 8.65% | **64.80%** | 40.63% | 23.56% | 8.42% | **72.60%** | 45.49% | 21.22% | 9.49% | **76.20%** |
| FSRM | 27.18% | 30.38% | 5.73% | **63.29%** | 39.81% | 22.39% | 9.91% | **72.11%** | 51.17% | 19.58% | 8.73% | **79.48%** | 56.20% | 17.79% | 10.19% | **84.18%** |
| **Moyenne** | **20.00%** | **22.91%** | **5.39%** | **48.31%** | **29.57%** | **17.97%** | **9.06%** | **56.60%** | **37.19%** | **17.08%** | **7.49%** | **61.75%** | **40.57%** | **16.05%** | **8.15%** | **64.78%** |

**MFCC Aurora - Variation vocabulaire**

| Nom Fichier | Configuration : modèle original, lw = 08, FillProb = 0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 15.12% | 12.07% | 6.63% | **33.82%** | 23.47% | 7.56% | 11.94% | **42.97%** | 31.17% | 7.16% | 13.79% | **52.12%** | 33.69% | 6.50% | 14.99% | **55.17%** |
| FDJM | 21.97% | 40.44% | 3.51% | **65.92%** | 32.07% | 34.13% | 7.25% | **73.45%** | 40.02% | 31.09% | 8.84% | **79.94%** | 42.31% | 30.34% | 8.93% | **81.58%** |
| FSRM | 28.22% | 26.06% | 5.87% | **60.14%** | 40.19% | 19.15% | 11.78% | **71.13%** | 53.99% | 13.66% | 12.68% | **80.33%** | 57.51% | 12.86% | 16.01% | **86.38%** |
| **Moyenne** | **21.77%** | **26.19%** | **5.34%** | **53.29%** | **31.91%** | **20.28%** | **10.32%** | **62.52%** | **41.73%** | **17.30%** | **11.77%** | **70.80%** | **44.50%** | **16.57%** | **13.31%** | **74.38%** |

TABLE B.11 – Résultats de la reconnaissance de parole sur le corpus domotique, eu utilisant les analyse acoustiques MFCC Sphinx&Aurora, en variant le vocabulaire et le modèle de langage associé.

4. Tableaux qui détaillent l'impact de quelques autres paramètres (poids du modèle de langage, probabilité des fillers) sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre, avec le vocabulaire domotique

**MFCC Sphinx - Variation poids LMw**

| Nom Fichier | Configuration : modèle original, vocab = 0.0k, FillProb = 0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 12.20% | 6.76% | 4.64% | **23.61%** | 11.54% | 5.84% | 3.98% | **21.35%** | 10.88% | 4.91% | 2.65% | **18.44%** | 11.94% | 3.98% | 2.79% | **18.70%** |
| FDJM | 20.62% | 31.60% | 5.80% | **58.02%** | 23.05% | 25.29% | 6.87% | **55.21%** | 24.73% | 21.27% | 7.20% | **53.20%** | 27.63% | 17.16% | 7.01% | **51.80%** |
| FSRM | 27.18% | 30.38% | 5.73% | **63.29%** | 28.69% | 26.06% | 6.15% | **60.89%** | 30.75% | 21.92% | 5.73% | **58.40%** | 33.24% | 17.56% | 6.38% | **57.18%** |
| **Moyenne** | **20.00%** | **22.91%** | **5.39%** | **48.31%** | **21.09%** | **19.06%** | **5.67%** | **45.82%** | **22.12%** | **16.03%** | **5.19%** | **43.35%** | **24.27%** | **12.90%** | **5.39%** | **42.56%** |

**MFCC Sphinx - Variation probabilites FillProb**

| Nom Fichier | Configuration : modèle original, vocab = 0.0k, lw = 05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FillProb=0.05 | | | | FillProb=0.01 | | | | FillProb=0.005 | | | | FillProb=0.001 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 11.94% | 3.98% | 2.79% | **18.70%** | 11.80% | 3.98% | 3.05% | **18.83%** | 11.94% | 3.98% | 3.05% | **18.97%** | 11.94% | 3.98% | 3.18% | **19.10%** |
| FDJM | 27.63% | 17.16% | 7.01% | **51.80%** | 27.07% | 16.64% | 7.67% | **51.38%** | 27.07% | 16.46% | 7.76% | **51.29%** | 26.60% | 15.99% | 7.85% | **50.44%** |
| FSRM | 33.24% | 17.56% | 6.38% | **57.18%** | 32.72% | 17.61% | 6.95% | **57.28%** | 32.91% | 17.42% | 7.28% | **57.61%** | 33.57% | 16.81% | 7.51% | **57.89%** |
| **Moyenne** | **24.27%** | **12.90%** | **5.39%** | **42.56%** | **23.86%** | **12.74%** | **5.89%** | **42.50%** | **23.97%** | **12.62%** | **6.03%** | **42.62%** | **24.04%** | **12.26%** | **6.18%** | **42.48%** |

**MFCC Aurora - Variation poids LMw**

| Nom Fichier | Configuration : modèle original, vocab = 0.0k, probaFiller = 0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 15.12% | 12.07% | 6.63% | **33.82%** | 16.45% | 10.34% | 6.10% | **32.89%** | 16.84% | 8.36% | 5.44% | **30.64%** | 19.23% | 6.23% | 9.15% | **34.62%** |
| FDJM | 21.97% | 40.44% | 3.51% | **65.92%** | 23.75% | 34.36% | 5.10% | **63.21%** | 26.23% | 29.27% | 6.31% | **61.80%** | 27.77% | 24.54% | 9.26% | **61.57%** |
| FSRM | 28.22% | 26.06% | 5.87% | **60.14%** | 30.52% | 22.21% | 7.46% | **60.19%** | 32.25% | 18.97% | 7.93% | **59.15%** | 34.13% | 16.38% | 8.92% | **59.44%** |
| **Moyenne** | **21.77%** | **26.19%** | **5.34%** | **53.29%** | **23.57%** | **22.30%** | **6.22%** | **52.10%** | **25.11%** | **18.87%** | **6.56%** | **50.53%** | **27.04%** | **15.72%** | **9.11%** | **51.88%** |

**MFCC Aurora - Variation probabilites FillProb**

| Nom Fichier | Configuration : modèle original, vocab = 0.0k, lw = 05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FillProb=0.05 | | | | FillProb=0.01 | | | | FillProb=0.005 | | | | FillProb=0.001 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 19.23% | 6.23% | 9.15% | **34.62%** | 19.23% | 6.37% | 9.81% | **35.41%** | 19.50% | 6.23% | 9.68% | **35.41%** | 19.10% | 6.23% | 9.81% | **35.15%** |
| FDJM | 27.77% | 24.54% | 9.26% | **61.57%** | 27.86% | 23.52% | 9.30% | **60.68%** | 27.54% | 23.84% | 9.68% | **61.06%** | 27.86% | 22.86% | 10.05% | **60.78%** |
| FSRM | 34.13% | 16.38% | 8.92% | **59.44%** | 34.65% | 15.59% | 9.34% | **59.58%** | 34.74% | 15.54% | 9.81% | **60.09%** | 34.79% | 15.26% | 10.38% | **60.42%** |
| **Moyenne** | **27.04%** | **15.72%** | **9.11%** | **51.88%** | **27.25%** | **15.16%** | **9.48%** | **51.89%** | **27.26%** | **15.20%** | **9.72%** | **52.19%** | **27.25%** | **15.78%** | **10.08%** | **52.12%** |

TABLE B.12 – Résultats de la reconnaissance de parole sur le corpus domotique, eu utilisant les analyse acoustiques MFCC Sphinx&Aurora, en variant le poids du modèle de langage et la probabilité des fillers.

5. Tableaux qui détaillent l'impact de l'adaptation des modèles acoustiques (MFCC Sphinx) sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre

**Modèle adapté - MLLR40cMAPmean**

| Nom Fichier | Configuration : modèle adapté MLLR40cMAPmean, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total |
| FDFM | 12.33% | 11.41% | 6.37% | **30.11%** | 21.88% | 8.22% | 10.74% | **40.85%** | 23.61% | 10.48% | 7.56% | **41.64%** | 23.61% | 10.88% | 6.10% | **40.58%** |
| FDJM | 18.98% | 32.40% | 6.26% | **57.64%** | 31.84% | 24.08% | 11.87% | **67.79%** | 43.38% | 21.37% | 11.83% | **76.58%** | 46.61% | 20.76% | 11.45% | **78.82%** |
| FSRM | 26.34% | 32.44% | 6.10% | **64.88%** | 39.58% | 23.10% | 9.67% | **72.53%** | 50.61% | 21.31% | 9.11% | **81.03%** | 55.54% | 18.69% | 9.44% | **83.66%** |
| **Moyenne** | **19.22%** | **25.42%** | **6.24%** | **50.88%** | **31.10%** | **18.47%** | **10.76%** | **60.39%** | **39.20%** | **17.72%** | **9.50%** | **66.42%** | **41.92%** | **16.78%** | **9.00%** | **67.69%** |

**Modèle adapté - MLLR40cMAP**

| Nom Fichier | Configuration : modèle adapté MLLR40cMAP, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total |
| FDFM | 13.00% | 11.54% | 5.31% | **29.84%** | 22.41% | 8.22% | 10.34% | **40.98%** | 24.40% | 10.48% | 7.43% | **42.31%** | 24.01% | 10.88% | 6.37% | **41.25%** |
| FDJM | 19.26% | 32.68% | 6.69% | **58.63%** | 31.32% | 23.38% | 11.08% | **65.78%** | 43.15% | 22.11% | 10.75% | **76.02%** | 45.72% | 21.65% | 9.58% | **76.95%** |
| FSRM | 25.21% | 32.68% | 6.57% | **64.46%** | 37.75% | 24.46% | 9.34% | **71.55%** | 50.42% | 21.36% | 7.75% | **79.53%** | 54.41% | 19.11% | 8.69% | **82.21%** |
| **Moyenne** | **19.16%** | **25.63%** | **6.19%** | **50.98%** | **30.49%** | **18.69%** | **10.25%** | **59.44%** | **39.32%** | **17.98%** | **8.64%** | **65.95%** | **41.38%** | **17.21%** | **8.21%** | **66.80%** |

**Modèle adapté - MLLR40c**

| Nom Fichier | Configuration : modèle adapté MLLR40c, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total |
| FDFM | 12.33% | 8.22% | 5.84% | **26.39%** | 20.56% | 7.16% | 9.68% | **37.40%** | 22.81% | 7.96% | 6.10% | **36.87%** | 23.34% | 9.02% | 5.31% | **37.67%** |
| FDJM | 18.42% | 23.84% | 8.13% | **50.40%** | 28.89% | 17.44% | 11.64% | **57.97%** | 36.93% | 17.02% | 10.43% | **64.38%** | 42.82% | 14.26% | 10.14% | **67.23%** |
| FSRM | 26.10% | 26.34% | 7.46% | **59.91%** | 38.45% | 18.54% | 12.02% | **69.01%** | 51.24% | 16.89% | 11.22% | **79.35%** | 54.76% | 15.49% | 11.36% | **81.60%** |
| **Moyenne** | **18.95%** | **19.47%** | **7.14%** | **45.57%** | **29.30%** | **14.38%** | **11.11%** | **54.79%** | **36.99%** | **13.96%** | **9.25%** | **60.20%** | **40.31%** | **12.92%** | **8.94%** | **62.17%** |

**Modèle adapté - MLLR1cMAPmean**

| Nom Fichier | Configuration : modèle adapté MLLR1cMAPmean, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total |
| FDFM | 12.20% | 12.60% | 5.97% | **30.77%** | 22.41% | 9.95% | 10.08% | **42.44%** | 24.14% | 11.67% | 8.36% | **44.16%** | 24.27% | 11.80% | 6.63% | **42.71%** |
| FDJM | 19.78% | 32.07% | 6.45% | **58.30%** | 32.91% | 23.42% | 12.11% | **68.44%** | 45.63% | 19.96% | 14.59% | **80.18%** | 47.69% | 19.96% | 11.92% | **79.57%** |
| FSRM | 25.07% | 33.33% | 6.10% | **64.51%** | 39.15% | 23.94% | 9.77% | **72.86%** | 52.96% | 20.52% | 8.92% | **82.39%** | 57,28% | 19.44% | 10.33% | **87.04%** |
| **Moyenne** | **19.02%** | **26.00%** | **6.17%** | **51.19%** | **31.49%** | **19.10%** | **10.65%** | **61.25%** | **40.91%** | **17.38%** | **10.62%** | **68.91%** | **35.98%** | **17.07%** | **9.63%** | **69.77%** |

**Modèle adapté - MLLR1cMAP**

| Nom Fichier | Configuration : modèle adapté MLLR1cMAP, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 12.33% | 11.14% | 6.10% | **29.58%** | 22.55% | 7.69% | 9.81% | **40.05%** | 23.87% | 10.61% | 6.37% | **40.85%** | 24.01% | 11.01% | 5.70% | **40.72%** |
| FDJM | 19.68% | 32.59% | 6.55% | **58.81%** | 31.56% | 23.56% | 11.45% | **66.57%** | 44.93% | 21.04% | 10.89% | **76.86%** | 47.69% | 20.57% | 9.77% | **78.03%** |
| FSRM | 25.45% | 32.35% | 6.29% | **63.99%** | 38.08% | 25.26% | 9.81% | **73.15%** | 50.19% | 21.41% | 8.36% | **79.95%** | 55.07% | 20.00% | 8.12% | **83.19%** |
| **Moyenne** | **19.15%** | **25.36%** | **6.31%** | **50.79%** | **30.73%** | **18.84%** | **10.36%** | **59.92%** | **39.66%** | **17.69%** | **8.54%** | **65.89%** | **42.26%** | **17.19%** | **7.86%** | **67.31%** |

**Modèle adapté - MLLR1c**

| Nom Fichier | Configuration : modèle adapté MLLR1c, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 11.14% | 8.36% | 6.10% | **25.60%** | 19.23% | 6.76% | 9.02% | **35.01%** | 21.75% | 8.36% | 4.64% | **34.75%** | 22.15% | 9.42% | 4.91% | **36.47%** |
| FDJM | 17.44% | 25.76% | 7.95% | **51.15%** | 28.80% | 18.56% | 10.89% | **58.25%** | 39.67% | 17.48% | 10.56% | **67.71%** | 44.27% | 16.04% | 10.38% | **70.69%** |
| FSRM | 24.32% | 27.46% | 6.90% | **58.69%** | 37.37% | 19.81% | 11.36% | **68.54%** | 50.99% | 16.71% | 10.09% | **77.79%** | 54.23% | 16.24% | 10.28% | **80.75%** |
| **Moyenne** | **17.63%** | **20.53%** | **6.98%** | **45.15%** | **28.47%** | **15.04%** | **10.42%** | **53.93%** | **37.47%** | **14.18%** | **8.43%** | **60.08%** | **40.22%** | **13.90%** | **8.52%** | **62.64%** |

**Modèle adapté - MAP**

| Nom Fichier | Configuration : modèle adapté MAP, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 12.47% | 10.48% | 5.70% | **28.65%** | 21.22% | 8.22% | 10.88% | **40.32%** | 22.15% | 10.88% | 5.97% | **38.99%** | 22.28% | 11.27% | 5.31% | **38.86%** |
| FDJM | 20.57% | 35.20% | 6.08% | **61.85%** | 31.09% | 27.72% | 10.14% | **68.96%** | 42.36% | 24.96% | 10.05% | **77.37%** | 45.91% | 24.17% | 9.58% | **79.66%** |
| FSRM | 25.77% | 35.77% | 5.68% | **67.23%** | 40.75% | 25.59% | 7.93% | **74.27%** | 51.88% | 23.05% | 7.75% | **82.68%** | 54.23% | 22.11% | 7.93% | **84.27%** |
| **Moyenne** | **19.60%** | **27.15%** | **5.82%** | **52.58%** | **31.02%** | **20.51%** | **9.65%** | **61.18%** | **38.80%** | **19.63%** | **7.92%** | **66.35%** | **40.81%** | **19.18%** | **7.61%** | **67.70%** |

**Modèle adapté - MAPmean**

| Nom Fichier | Configuration : modèle adapté MAPmean, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 12.60% | 10.74% | 6.37% | **29.71%** | 22.68% | 8.22% | 10.74% | **41.64%** | 24.01% | 10.74% | 8.89% | **43.63%** | 24.54% | 10.61% | 7.29% | **42.44%** |
| FDJM | 20.52% | 32.59% | 5.66% | **58.77%** | 33.66% | 23.94% | 11.36% | **68.96%** | 45.44% | 21.97% | 12.06% | **79.48%** | 50.49% | 19.59% | 12.62% | **82.70%** |
| FSRM | 25.07% | 33.43% | 5.63% | **64.13%** | 40.66% | 24.37% | 8.73% | **73.76%** | 50.09% | 22.35% | 9.01% | **81.46%** | 54.46% | 20.23% | 9.62% | **84.32%** |
| **Moyenne** | **19.40%** | **25.59%** | **5.89%** | **50.87%** | **32.33%** | **18.84%** | **10.28%** | **61.45%** | **39.45%** | **18.35%** | **9.99%** | **68.19%** | **43.16%** | **16.81%** | **9.84%** | **69.82%** |

TABLE B.13 – Résultats de la reconnaissance de parole sur le corpus domotique, eu utilisant l'analyse acoustique MFCC Sphinx, avec l'adaptation des modèles acoustiques.

6. Tableaux qui détaillent l'impact de l'adaptation des modèles acoustiques (MFCC Aurora) sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre

**Modèle adapté - MLLR40cMAPmean**

| Nom Fichier | Configuration : modèle adapté MLLR40cMAPmean, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 18.17% | 23.61% | 5.31% | **47.08%** | 31.17% | 16.58% | 11.14% | **58.89%** | 39.12% | 13.13% | 11.80% | **64.06%** | 41.25% | 11.94% | 12.60% | **65.78%** |
| FDJM | 17.81% | 56.24% | 1.92% | **75.97%** | 32.26% | 45.68% | 5.10% | **83.03%** | 39.83% | 42.45% | 5.84% | **88.13%** | 40.39% | 42.36% | 6.45% | **89.20%** |
| FSRM | 26.90% | 37.75% | 2.21% | **66.85%** | 48.50% | 22.77% | 8.50% | **79.77%** | 57,42% | 20.19% | 8.78% | **86.38%** | 60.19% | 19.15% | 11.08% | **90.42%** |
| **Moyenne** | **20.96%** | **39.20%** | **3.15%** | **63.30%** | **37.31%** | **28.34%** | **8.25%** | **73.90%** | **39.48%** | **25.26%** | **8.81%** | **79.52%** | **47.28%** | **24.48%** | **10.04%** | **81.80%** |

**Modèle adapté - MLLR40cMAP**

| Nom Fichier | Configuration : modèle adapté MLLR40cMAP, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 17.90% | 18.17% | 9.28% | **43.56%** | 29.84% | 13.40% | 12.20% | **55.44%** | 37.53% | 14.06% | 14.19% | 65.78% | 38.73% | 13.53% | 13.26% | **65.52%** |
| FDJM | 11.55% | 70.22% | 1.78% | **83.54%** | 17.91% | 67.41% | 3.88% | **89.20%** | 22.77% | 65.92% | 3.83% | 92.52% | 25.90% | 64.61% | 4.02% | **94.53%** |
| FSRM | 24.23% | 45.73% | 1.97% | **71.92%** | 42.07% | 34.27% | 4.04% | **80.38%** | 52.02% | 30.05% | 3.62% | 85.68% | 54.88% | 29.34% | 4.37% | **88.59%** |
| **Moyenne** | **17.89%** | **44.71%** | **4.34%** | **66.34%** | **29.94%** | **38.36%** | **6.71%** | **75.01%** | **37.44%** | **36.68%** | **7.21%** | **81.33%** | **39.84%** | **35.83%** | **7.22%** | **82.88%** |

**Modèle adapté - MLLR40c**

| Nom Fichier | Configuration : modèle adapté MLLR40c, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 16.31% | 14.06% | 8.22% | **38.59%** | 23.34% | 10.61% | 12.20% | **46.15%** | 31.17% | 10.21% | 11.01% | **52.39%** | 32.76% | 10.61% | 11.41% | **54.77%** |
| FDJM | 22.44% | 32.87% | 4.68% | **59.98%** | 34.36% | 24.96% | 8.93% | **68.26%** | 41.65% | 24.40% | 10.71% | **76.76%** | 46.19% | 23.05% | 10.89% | **80.13%** |
| FSRM | 29.95% | 28.17% | 4.41% | **62.54%** | 43.33% | 20.70% | 9.44% | **73.47%** | 55.26% | 16.10% | 10.47% | **81.83%** | 59.67% | 14.65% | 12.54% | **86.85%** |
| **Moyenne** | **22.90%** | **25.03%** | **5.77%** | **53.70%** | **33.68%** | **18.76%** | **10.19%** | **62.63%** | **42.69%** | **16.90%** | **10.73%** | **70.33%** | **46.21%** | **16.10%** | **11.61%** | **73.92%** |

**Modèle adapté - MLLR1cMAPmean**

| Nom Fichier | Configuration : modèle adapté MLLR1cMAPmean, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 19.10% | 22.81% | 6.10% | **48.01%** | 34.75% | 13.93% | 13.00% | **61.67%** | 39.12% | 13.40% | 16.31% | **68.83%** | 43.24% | 11.54% | 17.64% | **72.41%** |
| FDJM | 17.63% | 55.63% | 2.15% | **75.41%** | 32.54% | 44.46% | 5.52% | **82.52%** | 40.44% | 41.61% | 6.87% | **88.92%** | 41.33% | 41.61% | 7.15% | **90.09%** |
| FSRM | 26.95% | 37.56% | 2.63% | **67.14%** | 46.90% | 23.24% | 8.59% | **78.73%** | 56.15% | 20.52% | 9.67% | **86.34%** | 59.62% | 19.20% | 11.50% | **90.33%** |
| **Moyenne** | **21.23%** | **38.67%** | **3.63%** | **63.52%** | **38.06%** | **27.21%** | **9.04%** | **74.31%** | **45.24%** | **25.18%** | **10.95%** | **81.36%** | **48.06%** | **24.12%** | **12.10%** | **84.28%** |

**Modèle adapté - MLLR1cMAP**

| Nom Fichier | Configuration : modèle adapté MLLR1cMAP, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total |
| FDFM | 16.45% | 18.17% | 9.02% | **43.63%** | 26.66% | 14.19% | 11.80% | **52.65%** | 35.01% | 13.00% | 12.73% | **60.74%** | 35.81% | 12.86% | 12.33% | **61.01%** |
| FDJM | 10.38% | 72.46% | 1.64% | **84.48%** | 17.53% | 67.79% | 3.88% | **89.20%** | 22.25% | 67.23% | 3.74% | **93.22%** | 24.87% | 65.45% | 3.74% | **94.06%** |
| FSRM | 23.71% | 45.63% | 1.97% | **71.31%** | 42.07% | 34.13% | 3.90% | **80.09%** | 53.10% | 28.45% | 4.23% | **85.77%** | 55.40% | 27.75% | 5.21% | **88.36%** |
| **Moyenne** | **16.85%** | **45.42%** | **4.21%** | **66.47%** | **28.75%** | **38.70%** | **6.53%** | **73.98%** | **36.79%** | **36.23%** | **6.90%** | **79.91%** | **38.69%** | **35.35%** | **7.09%** | **81.14%** |

**Modèle adapté - MLLR1c**

| Nom Fichier | Configuration : modèle adapté MLLR1c, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total |
| FDFM | 16.05% | 13.66% | 7.56% | **37.27%** | 23.47% | 9.81% | 12.33% | **45.62%** | 31.43% | 9.95% | 12.47% | **53.85%** | 34.48% | 9.68% | 11.67% | **55.84%** |
| FDJM | 21.93% | 36.79% | 4.63% | **63.35%** | 32.07% | 29.03% | 7.53% | **68.63%** | 40.72% | 27.68% | 8.70% | **77.09%** | 44.69% | 25.57% | 10.52% | **80.79%** |
| FSRM | 26.90% | 29.25% | 4.65% | **60.80%** | 40.61% | 20.80% | 9.11% | **70.52%** | 51.08% | 17.98% | 8.78% | **77.84%** | 55.16% | 16.62% | 11.03% | **82.82%** |
| **Moyenne** | **21.63%** | **26.57%** | **5.61%** | **53.81%** | **32.05%** | **19.88%** | **9.66%** | **61.59%** | **41.08%** | **18.54%** | **9.98%** | **69.59%** | **44.78%** | **17.29%** | **11.07%** | **73.15%** |

**Modèle adapté - MAP**

| Nom Fichier | Configuration : modèle adapté MAP, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total |
| FDFM | 17.24% | 17.24% | 9.42% | **43.90%** | 28.12% | 13.00% | 12.60% | **53.71%** | 34.62% | 13.26% | 13.79% | **61.67%** | 37.27% | 11.94% | 13.53% | **62.73%** |
| FDJM | 10.71% | 71.95% | 2.10% | **84.76%** | 17.81% | 67.18% | 3.88% | **88.87%** | 21.74% | 67.74% | 3.79% | **93.27%** | 24.31% | 66.39% | 3.55% | **94.25%** |
| FSRM | 26.24% | 43.80% | 2.16% | **72.21%** | 43.85% | 31.83% | 5.16% | **80.85%** | 52.16% | 29.20% | 4.32% | **85.68%** | 55.21% | 26.90% | 4.55% | **86.67%** |
| **Moyenne** | **18.06%** | **44.33%** | **4.56%** | **66.96%** | **29.93%** | **37.34%** | **7.21%** | **74.48%** | **36.17%** | **36.73%** | **7.30%** | **80.21%** | **38.93%** | **35.08%** | **7.21%** | **81.22%** |

**Modèle adapté - MAPmean**

| Nom Fichier | Configuration : modèle adapté MAPmean, lw = 08, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | vocab = 0.0k | | | | vocab = 0.1k | | | | vocab = 0.5k | | | | vocab = 1k | | | |
| | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total | Err S | Err D | Err I | Total |
| FDFM | 19.50% | 21.22% | 6.23% | **46.95%** | 32.49% | 14.19% | 10.34% | **57.03%** | 38.06% | 13.26% | 13.13% | **64.46%** | 38.99% | 12.33% | 13.40% | **64.72%** |
| FDJM | 19.03% | 55.07% | 2.06% | **76.16%** | 33.85% | 43.29% | 6.08% | **83.22%** | 42.82% | 39.60% | 7.43% | **89.86%** | 44.13% | 39.41% | 8.32% | **91.87%** |
| FSRM | 27.32% | 35.59% | 3.47% | **66.38%** | 46.95% | 22.91% | 8.08% | **77.93%** | 59.11% | 17.70% | 9.91% | **86.71%** | 61.88% | 17.14% | 12.11% | **91.13%** |
| **Moyenne** | **21.95%** | **37.29%** | **3.92%** | **63.16%** | **37.76%** | **26.80%** | **8.17%** | **72.73%** | **46.66%** | **23.52%** | **10.16%** | **80.34%** | **48.33%** | **22.96%** | **11.28%** | **82.57%** |

TABLE B.14 – Résultats de la reconnaissance de parole sur le corpus domotique, eu utilisant l'analyse acoustique MFCC Aurora, avec l'adaptation des modèles acoustiques.

7. Tableaux qui détaillent l'impact du variation de poids du modèle de langage de différents vocabulaires sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre, avec le modèle adapté - MLLR40c

**Vocab 0.0k**

| Nom Fichier | Configuration : modèle adapté MLLR40c, voc = 0.0k, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 12.33% | 8.22% | 5.84% | **26.39%** | 13.13% | 7.16% | 4.64% | **24.93%** | 13.13% | 6.10% | 3.18% | **22.41%** | 13.13% | 5.17% | 2.79% | **21.09%** |
| FDJM | 18.42% | 23.84% | 8.13% | **50.40%** | 18.14% | 19.54% | 7.67% | **45.35%** | 18.47% | 17.02% | 7.29% | **42.78%** | 19.31% | 14.35% | 6.17% | **39.83%** |
| FSRM | 26.10% | 26.34% | 7.46% | **59.91%** | 26.06% | 23.33% | 7.98% | **57.37%** | 26.15% | 20.66% | 7.84% | **54.65%** | 26.81% | 18.26% | 7.09% | **52.16%** |
| **Moyenne** | **18.95%** | **19.47%** | **7.14%** | **45.57%** | **19.11%** | **16.68%** | **6.76%** | **42.55%** | **19.25%** | **14.59%** | **6.10%** | **39.95%** | **19.75%** | **12.59%** | **5.35%** | **37.69%** |

**Vocab 0.1k**

| Nom Fichier | Configuration : modèle adapté MLLR40c, voc = 0.1k, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 20.56% | 7.16% | 9.68% | **37.40%** | 20.29% | 6.37% | 9.81% | **36.47%** | 19.36% | 5.70% | 9.02% | **34.08%** | 19.89% | 5.31% | 8.49% | **33.69%** |
| FDJM | 28.89% | 17.44% | 11.64% | **57.97%** | 27.35% | 15.61% | 12.06% | **55.03%** | 28.61% | 13.32% | 12.58% | **54.51%** | 29.87% | 11.41% | 13.88% | **55.17%** |
| FSRM | 38.45% | 18.54% | 12.02% | **69.01%** | 37.79% | 17.28% | 13.76% | **68.83%** | 39.34% | 15.31% | 14.84% | **69.48%** | 39.48% | 13.43% | 15.77% | **68.69%** |
| **Moyenne** | **29.30%** | **14.38%** | **11.11%** | **54.79%** | **28.48%** | **13.09%** | **11.88%** | **53.44%** | **29.10%** | **11.44%** | **12.15%** | **52.69%** | **29.75%** | **10.05%** | **12.71%** | **52.52%** |

**Vocab 0.5k**

| Nom Fichier | Configuration : modèle adapté MLLR40c, voc = 0.5k, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 22.81% | 7.96% | 6.10% | **36.87%** | 22.15% | 7.69% | 5.84% | **35.68%** | 22.41% | 6.90% | 6.90% | **36.21%** | 22.55% | 5.70% | 8.22% | **36.47%** |
| FDJM | 36.93% | 17.02% | 10.43% | **64.38%** | 37.45% | 14.03% | 10.61% | **62.09%** | 37.66% | 12.43% | 11.45% | **61.54%** | 38.36% | 10.65% | 13.55% | **62.57%** |
| FSRM | 51.24% | 16.89% | 11.22% | **79.35%** | 50.16% | 15.39% | 11.87% | **77.43%** | 50.28% | 13.56% | 13.46% | **77.30%** | 50.66% | 11.96% | 16.04% | **78.66%** |
| **Moyenne** | **36.99%** | **13.96%** | **9.25%** | **60.20%** | **36.59%** | **12.37%** | **9.44%** | **58.40%** | **36.78%** | **10.96%** | **10.60%** | **58.35%** | **37.19%** | **9.44%** | **12.60%** | **59.23%** |

**Vocab 1k**

| Nom Fichier | Configuration : modèle adapté MLLR40c, voc = 1k, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 23.34% | 9.02% | 5.31% | **37.67%** | 22.55% | 7.96% | 5.04% | **35.54%** | 22.28% | 7.82% | 5.84% | **35.94%** | 22.94% | 6.23% | 7.29% | **36.47%** |
| FDJM | 42.82% | 14.26% | 10.14% | **67.23%** | 42.12% | 13.51% | 10.61% | **66.25%** | 41.37% | 11.50% | 11.97% | **64.84%** | 41.94% | 9.96% | 14.21% | **66.11%** |
| FSRM | 54.76% | 15.49% | 11.36% | **81.60%** | 55.42% | 13.19% | 12.58% | **81.18%** | 54.01% | 12.67% | 14.31% | **80.99%** | 54.69% | 11.26% | 17.96% | **83.91%** |
| **Moyenne** | **40.31%** | **12.92%** | **8.94%** | **62.17%** | **40.03%** | **11.55%** | **9.41%** | **60.99%** | **39.22%** | **10.66%** | **10.71%** | **60.59%** | **39.86%** | **9.15%** | **13.15%** | **62.16%** |

TABLE B.15 – Résultats de la reconnaissance de parole sur le corpus domotique, eu utilisant l'analyse acoustique MFCC Sphinx, en variant le vocabulaire et le poids du modèle de langage.

8. Tableaux qui détaillent l'impact du variation de poids du modèle de langage de différents vocabulaires sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre, avec le modèle adapté - MLLR40c

**Vocab 0.0k**

| Nom Fichier | Configuration : modèle adapté MLLR40c, voc = 0.0k, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 16.31% | 14.06% | 8.22% | **38.59%** | 16.71% | 11.67% | 6.90% | **35.28%** | 17.11% | 9.68% | 7.03% | **33.82%** | 17.11% | 8.49% | 6.76% | **32.36%** |
| FDJM | 22.44% | 32.87% | 4.68% | **59.98%** | 24.22% | 28.14% | 5.75% | **58.11%** | 27.40% | 23.89% | 7.62% | **58.91%** | 29.17% | 19.21% | 9.72% | **58.11%** |
| FSRM | 29.95% | 28.17% | 4.41% | **62.54%** | 30.56% | 25.54% | 5.02% | **61.13%** | 31.60% | 22.77% | 5.82% | **60.19%** | 33.43% | 19.53% | 7.42% | **60.38%** |
| **Moyenne** | **22.90%** | **25.03%** | **5.77%** | **53.70%** | **23.83%** | **21.78%** | **5.89%** | **51.51%** | **25.37%** | **18.78%** | **6.82%** | **50.97%** | **26.57%** | **15.74%** | **7.97%** | **50.28%** |

**Vocab 0.1k**

| Nom Fichier | Configuration : modèle adapté MLLR40c, voc = 0.1k, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 23.34% | 10.61% | 12.20% | **46.15%** | 23.34% | 9.15% | 13.26% | **45.76%** | 26.66% | 7.56% | 16.45% | **50.66%** | 27.45% | 6.37% | 17.11% | **50.93%** |
| FDJM | 34.36% | 24.96% | 8.93% | **68.26%** | 36.19% | 21.55% | 11.78% | **69.52%** | 38.76% | 18.33% | 14.12% | **71.20%** | 41.51% | 14.63% | 18.19% | **74.33%** |
| FSRM | 43.33% | 20.70% | 9.44% | **73.47%** | 44.37% | 17.65% | 12.11% | **74.13%** | 45.40% | 15.02% | 15.59% | **76.01%** | 47.32% | 12.58% | 19.15% | **79.06%** |
| **Moyenne** | **33.68%** | **18.76%** | **10.19%** | **62.63%** | **34.63%** | **16.12%** | **12.38%** | **63.14%** | **36.94%** | **13.64%** | **15.39%** | **65.96%** | **38.76%** | **11.19%** | **18.15%** | **68.11%** |

**Vocab 0.5k**

| Nom Fichier | Configuration : modèle adapté MLLR40c, voc = 0.5k, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 31.17% | 10.21% | 11.01% | **52.39%** | 33.02% | 8.36% | 12.73% | **54.11%** | 33.82% | 6.63% | 17.51% | **57.96%** | 34.08% | 6.10% | 22.28% | **62.47%** |
| FDJM | 41.65% | 24.40% | 10.71% | **76.76%** | 44.23% | 20.01% | 12.34% | **76.58%** | 47.27% | 16.41% | 14.59% | **78.26%** | 49.79% | 12.39% | 18.65% | **80.83%** |
| FSRM | 55.26% | 16.10% | 10.47% | **81.83%** | 56.90% | 13.38% | 13.15% | **83.43%** | 57.79% | 11.13% | 19.62% | **88.54%** | 58.83% | 8.87% | 24.13% | **91.83%** |
| **Moyenne** | **42.69%** | **16.90%** | **10.73%** | **70.33%** | **44.72%** | **13.92%** | **12.74%** | **71.37%** | **46.29%** | **11.39%** | **17.24%** | **74.92%** | **47.57%** | **9.12%** | **21.69%** | **78.38%** |

**Vocab 1k**

| Nom Fichier | Configuration : modèle adapté MLLR40c, voc = 1k, FillProb=0.05 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lw=08 | | | | lw=07 | | | | lw=06 | | | | lw=05 | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 32.76% | 10.61% | 11.41% | **54.77%** | 34.75% | 8.36% | 13.53% | **56.63%** | 35.68% | 7.29% | 16.31% | **59.28%** | 36.07% | 6.10% | 21.09% | **63.26%** |
| FDJM | 46.19% | 23.05% | 10.89% | **80.13%** | 48.76% | 18.37% | 13.18% | **80.32%** | 49.65% | 15.80% | 15.94% | **81.39%** | 51.99% | 12.44% | 19.73% | **84.15%** |
| FSRM | 59.67% | 14.65% | 12.54% | **86.85%** | 59.81% | 12.54% | 15.87% | **88.22%** | 61.08% | 10.38% | 20.19% | **91.64%** | 61.64% | 8.54% | 26.10% | **96.29%** |
| **Moyenne** | **46.21%** | **16.10%** | **11.61%** | **73.29%** | **47.77%** | **13.09%** | **14.19%** | **75.06%** | **48.80%** | **11.16%** | **17.48%** | **77.44%** | **49.90%** | **9.03%** | **22.31%** | **81.23%** |

TABLE B.16 – Résultats de la reconnaissance de parole sur le corpus domotique, eu utilisant l'analyse acoustique MFCC Aurora, en variant le vocabulaire et le poids du modèle de langage.

9. Tableau qui détaille les meilleurs résultats de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre, en utilisant l'analyse acoustique MFCC Sphinx

| **Nom Fichier** | **Configuration : modèle adapté MLLR40c, voc = 0.0k, lw=05** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **FillProb=0.05** | | | | **FillProb=0.01** | | | | **FillProb=0.005** | | | | **FillProb=0.001** | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 13.13% | 5.17% | 2.79% | **21.09%** | 13.13% | 5.31% | 2.79% | **21.22%** | 13.13% | 5.31% | 2.79% | **21.22%** | 13.13% | 5.31% | 2.92% | **21.35%** |
| FDJM | 19.31% | 14.35% | 6.17% | **39.83%** | 19.54% | 13.79% | 6.50% | **39.83%** | 19.50% | 13.79% | 6.40% | **39.69%** | 19.54% | 13.46% | 6.55% | **39.55%** |
| FSRM | 26.81% | 18.26% | 7.09% | **52.16%** | 26.85% | 18.17% | 7.04% | **52.07%** | 26.67% | 18.36% | 7.14% | **52.16%** | 26.71% | 18.17% | 7.32% | **52.21%** |
| **Moyenne** | **19.75%** | **12.59%** | **5.35%** | **37.69%** | **19.84%** | **12.42%** | **5.44%** | **37.71%** | **19.77%** | **12.49%** | **5.44%** | **37.69%** | **19.79%** | **12.31%** | **5.60%** | **37.70%** |

TABLE B.17 – Meilleurs résultats de la reconnaissance de parole sur le corpus domotique, en utilisant l'analyse acoustique MFCC Sphinx

10. Tableau qui détaille les meilleurs résultats de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre, en utilisant l'analyse acoustique MFCC Aurora

| **Nom Fichier** | **Configuration : modèle adapté MLLR40c, voc = 0.0k, lw=05** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **FillProb=0.05** | | | | **FillProb=0.01** | | | | **FillProb=0.005** | | | | **FillProb=0.001** | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 17.11% | 8.49% | 6.76% | **32.36%** | 17.51% | 7.69% | 6.90% | **32.10%** | 17.51% | 7.69% | 6.90% | **32.10%** | 17.77% | 7.29% | 7.03% | **32.10%** |
| FDJM | 29.17% | 19.21% | 9.72% | **58.11%** | 29.31% | 18.61% | 10.99% | **58.91%** | 29.13% | 18.56% | 11.08% | **58.77%** | 29.59% | 17.30% | 11.73% | **58.63%** |
| FSRM | 33.43% | 19.53% | 7.42% | **60.38%** | 33.85% | 17.84% | 8.50% | **60.19%** | 33.76% | 17.75% | 8.87% | **60.38%** | 34.04% | 17.28% | 10.00% | **61.31%** |
| **Moyenne** | **26.57%** | **15.74%** | **7.97%** | **50.28%** | **26.89%** | **14.71%** | **8.80%** | **50.40%** | **26.80%** | **14.67%** | **8.95%** | **50.42%** | **27.13%** | **13.96%** | **9.59%** | **50.68%** |

TABLE B.18 – Meilleurs résultats de la reconnaissance de parole sur le corpus domotique, en utilisant l'analyse acoustique MFCC Aurora.

11. Tableau qui détaille les meilleurs résultats de la reconnaissance de parole sur le corpus domotique (segments commande) enregistré en continu à une distance de 1 mètre, en utilisant l'analyse acoustique MFCC Sphinx

| **Nom Fichier** | **Configuration : modèle adapté MLLR40c, voc = 0.0k, lw=05** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **FillProb=0.05** | | | | **FillProb=0.01** | | | | **FillProb=0.005** | | | | **FillProb=0.001** | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 10.14% | 3.77% | 2.36% | **16.27%** | 9.91% | 4.01% | 2.36% | **16.27%** | 9.91% | 4.01% | 2.36% | **16.27%** | 9.91% | 4.01% | 2.59% | **16.51%** |
| FDJM | 17.15% | 14.42% | 7.56% | **39.14%** | 17.33% | 13.72% | 8.18% | **39.23%** | 17.24% | 13.72% | 8.18% | **39.14%** | 17.41% | 13.46% | 8.27% | **39.14%** |
| FSRM | 25.13% | 17.71% | 9.08% | **51.92%** | 25.31% | 17.45% | 8.99% | **51.75%** | 24.69% | 17.89% | 9.25% | **51.83%** | 24.61% | 17.80% | 9.42% | **51.83%** |
| **Moyenne** | **17.47%** | **11.97%** | **6.33%** | **35.78%** | **17.52%** | **11.73%** | **6.51%** | **35.75%** | **17.28%** | **11.87%** | **6.60%** | **35.75%** | **17.31%** | **11.76%** | **6.76%** | **35.83%** |

TABLE B.19 – Meilleurs résultats de la reconnaissance de parole sur le corpus domotique (segments commande), en utilisant l'analyse acoustique MFCC Sphinx

12. Tableau qui détaille les meilleurs résultats de la reconnaissance de parole sur le corpus domotique (segments non-commande) enregistré en continu à une distance de 1 mètre, en utilisant l'analyse acoustique MFCC Sphinx

| **Nom Fichier** | **Configuration : modèle adapté MLLR40c, voc = 0.0k, lw=05** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **FillProb=0.05** | | | | **FillProb=0.01** | | | | **FillProb=0.005** | | | | **FillProb=0.001** | | | |
| | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 16.97% | 6.97% | 3.33% | **27.27%** | 17.27% | 6.97% | 3.33% | **27.58%** | 17.27% | 6.97% | 3.33% | **27.58%** | 17.27% | 6.97% | 3.33% | **27.58%** |
| FDJM | 21.76% | 14.17% | 4.69% | **40.62%** | 22.06% | 13.97% | 4.49% | **40.52%** | 22.06% | 13.87% | 4.39% | **40.32%** | 21.86% | 13.57% | 4.49% | **39.92%** |
| FSRM | 28.76% | 18.90% | 4.88% | **52.54%** | 28.66% | 19.00% | 4.78% | **52.44%** | 28.96% | 18.90% | 4.78% | **52.64%** | 29.17% | 18.60% | 4.88% | **52.64%** |
| **Moyenne** | **22.50%** | **13.35%** | **4.30%** | **40.14%** | **22.66%** | **13.31%** | **4.20%** | **40.18%** | **22.76%** | **13.25%** | **4.17%** | **40.18%** | **22.77%** | **13.05%** | **4.23%** | **40.05%** |

TABLE B.20 – Meilleurs résultats de la reconnaissance de parole sur le corpus domotique (segments non-commande), en utilisant l'analyse acoustique MFCC Sphinx.

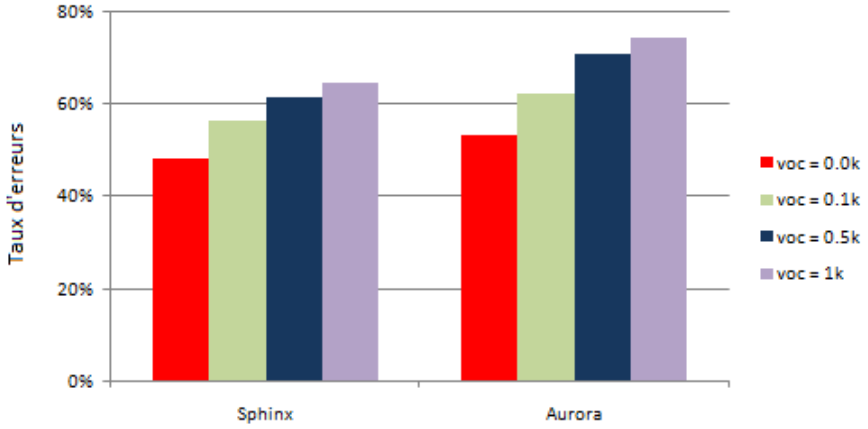
13. Tableau qui détaille les erreur de reconnaissance pour le mot « Majordome » dans le corpus domotique (segments commande) enregistré en continu à une distance de 1 mètre, en utilisant l'analyse acoustique MFCC Sphinx

| **Nom Fichier** | **Nb. occurences « Majordome »** | **Configuration - MLLR40c, 0.0k, 05, 0.005** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Nb. erreur reconnaissance** | | | | **Pourcentage erreur reconnaissance** | | | |
| | | Err S | Err D | Err I | **Total** | Err S | Err D | Err I | **Total** |
| FDFM | 50 | 0 | 0 | 0 | 0 | 0.00% | 0.00% | 0.00% | 0.00% |
| FDJM | 142 | 8 | 2 | 2 | 12 | 5,63% | 1.39% | 1.39% | **8.45%** |
| FSRM | 144 | 28 | 2 | 0 | 30 | 19.44% | 1.38% | 0.00% | **20.83%** |
| **Somme / Moyenne** | **336** | **36** | **4** | **2** | **42** | **10.71%** | **1.19%** | **0.59%** | **12.50%** |

TABLE B.21 – Analyse des erreur de reconnaissance pour le mot « Majordome » dans le corpus domotique (segments commande), en utilisant l'analyse acoustique MFCC Sphinx.

1. Graphiques qui détaillent

– l'impact du vocabulaire et du modèle de langage associé sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre. On compare les résultats obtenus pour les analyses acoustiques MFFC Sphinx & Aurora.




– l'impact de quelques autres paramètres (poids du modèle de langage, probabilité des fillers) sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre, avec le vocabulaire domotique. On compare les résultats obtenus pour les analyses acoustiques MFFC Sphinx & Aurora.

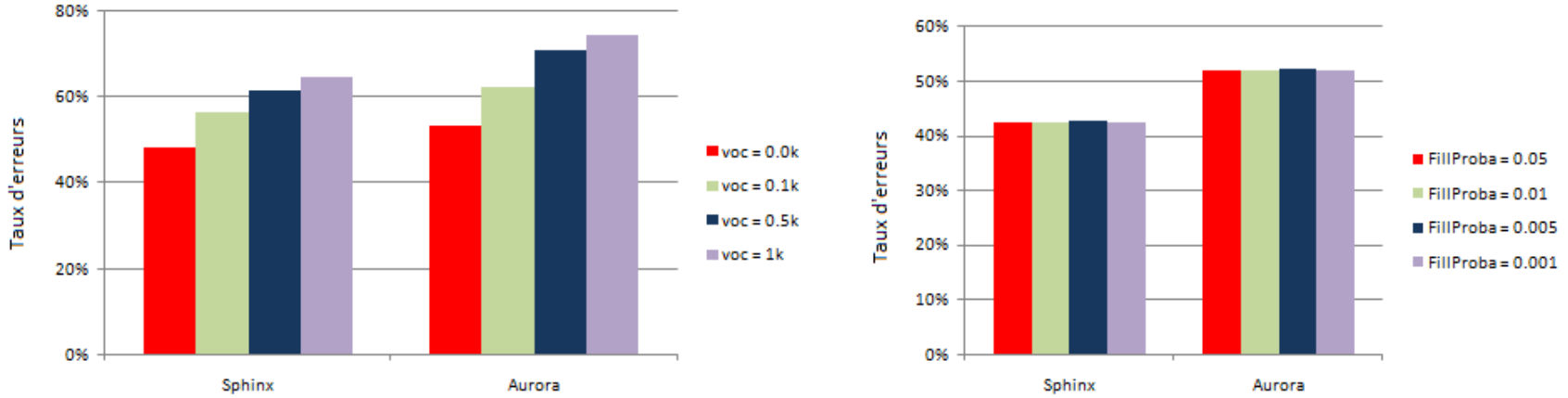



FIGURE B.1 – Résultats avec analyse MFCC Sphinx & Aurora (cf. tableaux B.11 et B.12).

2. Graphique qui détaille l'impact de l'adaptation des modèles acoustiques sur les performances de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre. On compare les résultats obtenus pour les analyses acoustiques MFFC Sphinx & Aurora.

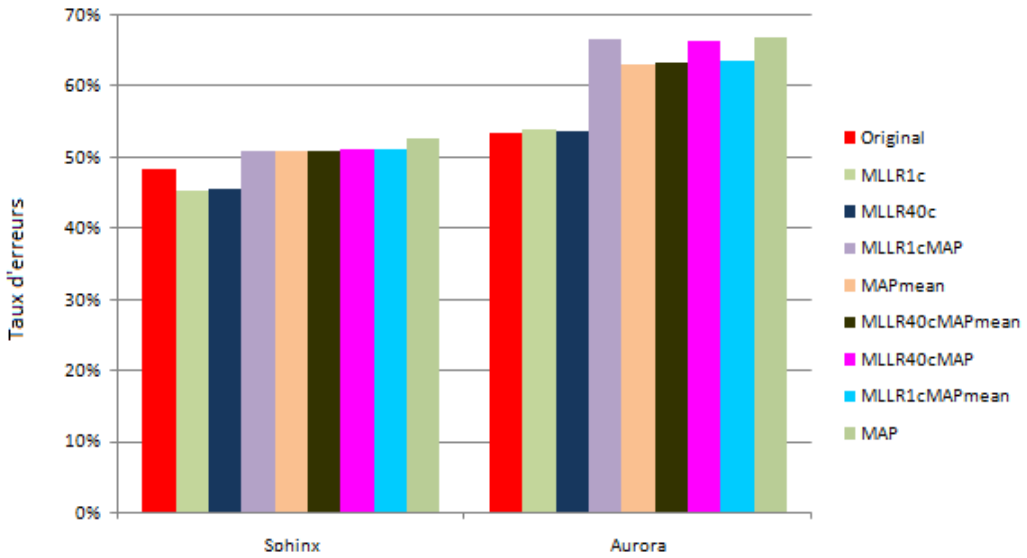



FIGURE B.2 – Résultats avec analyse MFCC Sphinx & Aurora (cf. tableau B.13 et B.14).

3. Graphiques qui détaillent les meilleurs résultats de la reconnaissance de parole sur le corpus domotique enregistré en continu à une distance de 1 mètre. On compare les résultats obtenus pour les analyses acoustiques MFFC Sphinx & Aurora.

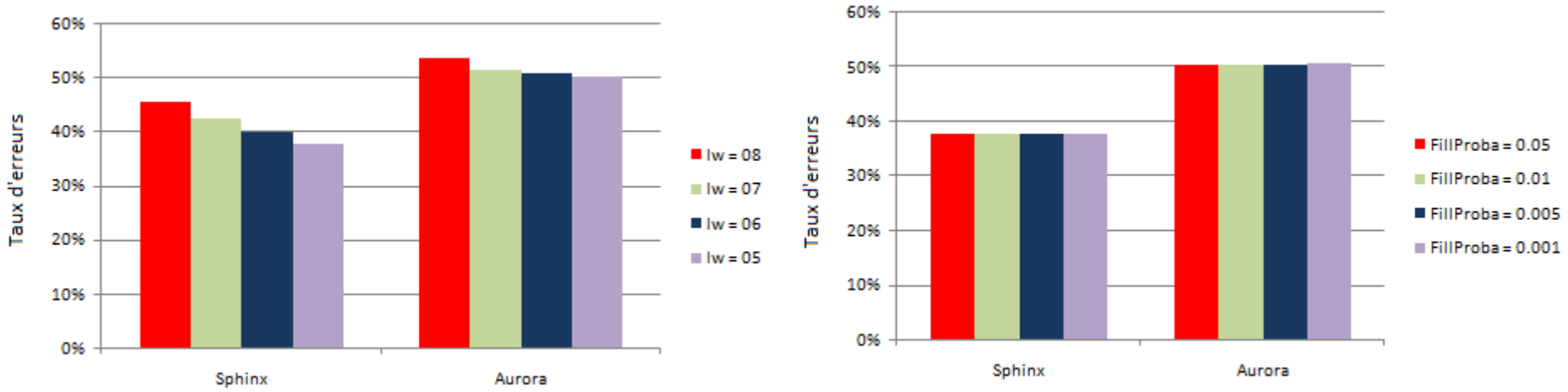



FIGURE B.3 – Résultats avec analyse MFCC Sphinx & Aurora (cf. tableaux B.15, B.16, B.17 et B.18).

## B.3 Résultats sur corpus CHIME

Tableau détaillant les résultats de la reconnaissance de parole sur le corpus CHIME, en utilisant l'-analyse acoustique MFCC Aurora

| **Type paramétrage** | **SNR** | | | | | |
|---|---|---|---|---|---|---|
| | **-6dB** | **-3dB** | **0dB** | **3dB** | **6dB** | **9dB** |
| MFCC HTK | 68.92% | 63.25% | 50.92% | 36.00% | 26.17% | 16.92% |
| MFCC Aurora | 68.08% | 60.08% | 48.00% | 35.92% | 24.00% | 16.42% |

TABLE B.22 – Résultats de la reconnaissance de parole sur le corpus CHIME, en utilisant les analyse acoustiques MFCC HTK & Aurora.

# Annexe C

# Plus de détails sur les outils utilisés

Cette annexe fournit quelques informations sur les outils qu'on a utilisé dans notre analyse.

## C.1 Paramétrisation

– **AcousticAnalysisDirectory.pl**
$Objet$ : analyse acoustique des fichiers signaux (wav) d'un répertoire.
$Utilisation$ : AcousticAnalysisDirectory.pl [-help] -type <analyse> -iDir <dir> [-iExt <ext>] -oDir <dir> [-oExt <ext>] [-tmpDir <dir>] [-debug]
$Options$ :
-help : affichage de l'aide
-iDir <dir> : répertoire contenant fichiers signaux
-iExt <ext> : extension des fichiers signal (défaut = $wav$)
-oDir <dir> : répertoire contenant fichiers analyses
-oExt <ext> : extension des fichiers analyses (défaut = $mfcc$)
-tmpDir <dir> : répertoire pour fichiers temporaires (défaut = ./)
-debug : affichage d'informations lors du traitement.
-type <analyse> : types analyses acoustiques :
– 16k_MFCC_Standard_Sphinx <=> signal 16 kHz, analyse MFCC standard => 13 coefs.
– 16k_Aurora_13MFCC_Sphinx <=> signal 16 kHz, analyse MFCC => 13 coefs.
– 16k_MFCC_E_Htk <=> signal 16 kHz, analyse MFCC standard => 13 coefs.

– **ProcessingFeatures**
$Objet$ : extrait les informations gardes dans un fichier analysé ($.mfcc$).
$Utilisation$ : ProcessingFeatures -srcFn <srcFile> -srcFmt <HTK> [-srcFrameCoeffs <nbCoeff-PerFrame>] [-srcFramePeriod <HTKFramePeriod>] [-select <i1,i2,i3,...,iN>] [-addDerivatives <d1Wsize> <d2Wsize>] [-display <firstframe> <lastframe>] [-displayLine <nbcoeff>] [-displayStats] [-dstFn <dstFile> -dstFmt <Sphinx>]

$Options$ :
-srcFn <srcFile> : spécifie le nom du fichier d'entrée
-srcFmt <format> : spécifie le format du fichier d'entrée (HTK ou Sphinx)
-srcFrameCoeffs <nbCoeffPerFrame> : spécifie le nombre du coefficients per trame (obligatoire si le format est Sphinx)
-srcFramePeriod <HTKFramePeriod> : la période de trame HTK (obligatoire si le format d'entrée est Sphinx, et le format de sortie - HTK)
-display <first> <last> : donne la liste des coefficients, a partir du trame <first> jusqu'à trame <last> (le premier trame est 0) ; si lastFrame = -1, tout le fichier va être transformé
-dstFn <dstFile> : spécifie le nom du fichier de sortie
-dstFmt <format> : spécifie le format du fichier de sortie (défaut : Sphinx).

– **processDirFeatures.sh**
$Objet$ : scripte qui applique l'outil $ProcessingFeatures$ sur chaque fichier d'un répertoire.
$Utilisation$ : processDirFeatures.sh <iDir> <oDir>
$Options$ :
-iDir : indique le répertoire qui contient les fichiers analyses
-oDir : indique le répertoire pour sauvegarder les résultats.

**Exemple de parametrisation :**
AcousticAnalysisDirectory.pl \
-iDir /home/luiza/Prog/git/Files/AudioFiles/Original/Audio_Dev \
-oDir /home/luiza/Prog/git/Files/AudioFilesAnalyzed/audio_dev/ -type 16k_MFCC_Standard_Sphinx

processDirFeatures.sh \
/home/luiza/Prog/git/Files/AudioFilesAnalyzed/audio_dev \
/home/luiza/Prog/git/Files/txtFiles/txtAudioFiles/audio_dev

## C.2 Fichiers de transcriptions alignées

– L'alignement entre fichiers audio - **match.jar**
$Objet$ : calcule le chemin optimal entre les fichiers originaux et leur correspondants enregistrés.
$Utilisation$ : java -jar match.jar Directories <srcDirectory1> <srcDirectory2> <srcDirectoryOutput> [ <format> ]
ou
Files <srcFn1> <srcFn2> <srcDirectoryOutput> [ <format> ] [ <size> ]
$Options$ :
-srcDirectory1 : spécifie le répertoire d'entrée (avec les fichiers originaux analyses)
-srcDirectory2 : spécifie le répertoire d'entrée (avec les fichiers enregistrés analyses)
-srcDirectoryOutput : spécifie le répertoire de sortie
-srcFn1 : spécifie le fichier d'entrée (original)
-srcFn2 : spécifie le fichier d'entrée (enregistré)
-format : spécifie le format du fichiers d'entrée : HTK ou Sphinx

– Création des fichiers .stm (alignés par DP) - **CheckAlignment.jar**
$Objet$ : donne les fichiers de transcriptions alignées (en utilisant les chemins optimaux obtenus par DP et les fichiers originaux de transcriptions).
$Utilisation$ : java -jar CheckAlignment.jar Directories <stmDir> <pathDir> <outputDir>
ou
Files <stmFn> <pathFn> <outputDir>
$Options$ :
-stmFn : spécifie le fichier qui garde la transcription du fichier original ($.stm$)
-pathFn : spécifie le fichier qui garde le chemin optimal entre un fichier original et son enregistrement ($.txt$)
-stmDir : spécifie le répertoire qui garde les transcriptions des fichiers originaux ($.stm$)
-pathDir : spécifie le répertoire qui garde les chemin optimaux entre les fichiers originaux et leurs enregistrements ($.txt$)
-outputDir : spécifie le répertoire de sortie qui va contenir les transcriptions alignées ($.stm$).

– Création des fichiers .stm (réalignés par régression linéaire) - **Align.pl**
$Objet$ : trouve la ligne la meilleure qui passe parmi un ensemble des points 2D.
$Utilisation$ : Align.pl [-help] [-files <fileNameList>] [-ignore <nbFrames>] [-cumul <pourcentage>] [-debug]
$Options$ :
-help : affichage de l'aide.
-files <filenameList> : précise la liste des fichiers (qui gardent les chemins optimaux entre les fichiers originaux et leurs enregistrements) qui vont être transformés
-ignore <nbFrames> : précise le nombre de trames qui vont être ignorés au début et à la fin du fichier
-cumul <pourcentage> : pourcentage cumulé à atteindre pour la sélection.

- java -jar **GenerateSTMs**.jar
$Objet$ : donne les fichiers de transcriptions réalignés (en utilisant les lignes obtenus par régression linéaire et les fichiers originaux de transcriptions)
$Utilisation$ : <logFile> <stmDirectory> <outputDirectory>
$Options$ :
-logFile : précise le fichier log obtenu après Align.pl
-stmDir : spécifie le répertoire qui garde les transcriptions des fichiers originaux ($.stm$)
-outputDir : spécifie le répertoire de sortie qui va contenir les transcriptions réalignées ($.stm$).

## C.3 Adaptation des modèles

– **forAdapt.jar**

$Objet$ : utilisé dans le cas de l'adaptation non-supervisée, pour obtenir, à partir de fichiers $.res$ résultants de la transcription automatique, les fichiers $.ctl$ et $.trans$ nécessaires pour l'adaptation.
$Utilisation$ : java -jar forAdapt.jar <resDir> <outputDir>
$Options$ :
-<resDir> : le répertoire où se trouvent les résultats de la transcription (on s'intéresse juste aux fichiers $.res$)
-<outputDir> : le répertoire pour sauvegarder les résultats (fichiers $.ctl$ et $.trans$ précisant les données et transcriptions pour adaptation).

– **sed**

$Objet$ : modification des annotations des bruits dans le fichier d'entrée, en suivant les règles d'un script.
$Utilisation$ : sed -f <script-file> <input-file>
$Options$ :
-f <script-file> : ajoute le contenu du <script-file> pour les commandes à exécuter
-<input-file> : précise le fichier à modifier.

– **normalize**

$Objet$ : normalise les fichiers au format texte, xml ou stm.
$Utilisation$ : normalize [-if input_format] [-of output_format] [-n] [-l] [-ie input_encoding] [-oe output_encoding] [-i input_file] [-o output_file]
$Options$ :
-if input_format : spécifie le format d'entrée 'input_format'
-of output_format : spécifie le format de sortie 'output_format '
-n : normalise le texte d'entrée en enlevant les espaces multiples, signes de ponctuation, et autres marques (comme '-')
-l : converti les chiffres en lettres (en français).
-ie input_encoding : spécifie l'encodage d'entrée 'input_encoding'
-oe output_encoding : spécifie l'encodage de sortie 'output_encoding'
-i input_file : spécifie le fichier d'entrée
-o output_file : spécifie le fichier de sortie.

– **iconv**

$Objet$ : Converti l'encodage des fichiers d'un encodage à un autre

$Utilisation$ : iconv -f <inputFileFormat> -t <outputFileFormat> [-l] [-c] [-s] <inputFile> -o <outputFile>.

$Options$ :

-f : spécifie l'encodage d'entrée

-t : spécifie l'encodage de sortie

-l : donne la liste de tous les encodages connus

-c : omet les caractères non valides de la sortie

-<inputFile> : spécifie le fichier d'entrée

-o : spécifie le fichier de sortie

-s : supprime les avertissements.

– **SphinxSelectSegments.pl**

$Objet$ : sélection (en fonction du lexique disponible) d'un sous ensemble des segments du corpus en vue de l'apprentissage de modèles acoustiques avec Sphinx.

$Utilisation$ : SphinxSelectSegments.pl -help -stm <fileName> -ctl <fileName> -trans <fileName> [-wrd <fileName>] [-lex <fileName>] [-oovFile <fileName>]

$Options$ :

-stm <fileName> : fichier contenant les segments disponibles

-radio <re radio> : expression régulière pour sélection selon la radio

-ctl <fileName> : liste des segments a utiliser pour l'apprentissage de modèles acoustiques avec Sphinx

-trans <fileName> : fichier contenant les transcriptions correspondant aux segments

-wrd <fileName> : permet de préciser un nom de fichier pour l'écriture des statistiques sur les occurrences des mots dans corpus sélectionné (sortie)

-lex <fileName> : permet de préciser un fichier lexique (les segments contenant des mots hors de ce lexique ne sont pas conservés)

-oovFile <fileName> : spécifie le nom de fichier pour l'écriture de la liste des mots hors vocabulaire (sortie).

– **SphinxAlignSpeechSegments.pl**
$Objet$ : réalise l'alignement sur les données de parole (pour déterminer quelles variantes de prononciation ont été utilisées).
$Utilisation$ : SphinxAlignSpeechSegments.pl -mdef <file> -hmm <dir> -dict <file> -fdict <file> -cepDir <dir> -cepExt <ext> -inCtl <file> -inTrans <file> -outCtl <file> -outTrans <file>
$Options$ :
-mdef <file> : spécifie le fichier comprenant la définition du modèle
-hmm <dir> : spécifie le répertoire comprenant les paramètres du modèle HMM (mean, variance, weights, transition matrix)
-dict <file> : spécifie le dictionnaire pour silences & bruits
-fdict <file> : spécifie les données de remplissage (fillers)
-cepDir <dir> : spécifie le répertoire contenant les fichiers analysés
-cepExt <ext> : spécifie l'extension pour fichiers analysés (défaut = .mfcc)
-inCtl <file> : fichier d'entré, précisant les données a traiter
-inTrans <file> : fichier d'entré synchrone de "control file" pour transcriptions
-outCtl <file> : fichier "control file" de sortie, correspondant au données bien alignées (sortie)
-outTrans <file> : fichier de transcriptions de sortie, correspondant au données bien alignées (sortie).

– **SphinxCreateAndTrainAcousticModels.pl**
$Objet$ : fabrication, apprentissage ou adaptation des modèles acoustiques
*Options pour adaptation des modèles* :
-inMdefFn <file> : description du modèle (.mdef)
-inMdlParamsDir <dir> : paramètres du modèle (avant adaptation)
-outMdlParamsDir <dir> : paramètres du modèle (après adaptation) (sortie)
-adaptation <typeAdapt> : mode d'adaptation : MAP (MAPIter=mean,var,tmat,mixw - adaptation des parametrs indiques uniquement)
-dictFn <file> : dictionnaire des prononciations des mots
-fdictFn <file> : dictionnaire pour silences & bruits
-ctlFn <file> : fichier « control file » précisant les données a traiter
-transFn <file> : fichier synchrone de « control file » pour transcriptions
-cepDir <dir> : répertoire contenant les fichiers analysés
-cepExt <.ext> : extension pour fichiers analysés (défaut = .mfcc)
-featType <type> : permet de spécifier type de paramétrage (défaut = 1s_c_d_dd => calcul online des dérivées)
-featLength <nbCoeff> : nombre de coefficients par trame (défaut = 13)
-cmn <no/current> : Cepstram Mean Normalisation (défaut = current)
-varNorm <no/yes> : normalisation des variances par enregistrement/segment (défaut = no)
-varFloor <value> : seuil sur variances pour Baum-Welch computations
-maxIter <nb> : nombre max d'itérations d'apprentissage (défaut = 20)
-tau <valeur> : précise la valeur de $\tau$ pour adaptation bayesienne (MAP ou MAPIter)
-bayesMean <yes/no> : précise la valeur de cette option pour adaptation (défaut : "no" si option -tau précisée, "yes" sinon).

– **bw**

$Objet$ : implémentation de l'algorithme Baulm-Welch pour recueillir les statistiques nécessaires à l'estimation des paramètres des modèles acoustiques.

$Utilisation$ : bw -ts2cbfn .cont. -moddeffn <.mdef> -mixwfn </mixture_weights> -meanfn </means> -varfn </variances> -tmatfn </transition_matrices> -dictfn <.lex> -fdictfn <.filler> -cepdir <cepDir> -cepext mfcc -lsnfn <.trans> -meanreest yes -varreest no -2passvar yes -feat 1s_c_d_dd -ceplen 13 -ctlfn <.ctl> -accumdir </AccumDir> -agc none -cmn current

$Options$ :

-moddeffn : description du modèle (.mdef)

-ts2cbfn .cont.

-mixwfn : spécifie les « mix weights » du modèle HMM

-tmatfn : spécifie la matrice de transitions du modèle HMM

-meanfn : spécifie les « means » du modèle HMM

-varfn : spécifie les « variances » du modèle HMM

-dictfn : dictionnaire des prononciations des mots

-fdictfn : dictionnaire pour silences & bruits

-ctlfn : fichier « control file » précisant les données a traiter

-cepdir : répertoire contenant les fichiers analysés

-cepext : extension pour fichiers analysés (défaut = .mfcc)

-lsnfn : fichier synchrone de « control file » pour transcriptions

-accumdir : répertoire pour accumuler les moyennes (sortie)

-meanreest : ré-estimation de la moyenne (yes/no)

-varreest : ré-estimation de la variance (yes/no)

-tmatreest : ré-estimation de la matrice de transition (yes/no)

-feat : permet de spécifier le type de paramétrage (défaut = 1s_c_d_dd => calcul online des dérivées)

-ceplen : nombre de coefficients par trame (défaut = 13)

-cmn : Cepstram Mean Normalisation (défaut = current).

– **mllr_solve**

$Objet$ : calcule la matrice de transformation linéaire basée sur le critère ML.

$Utilisation$ : mllr_solve -outmllrfn <adapt.matrix> -accumdir </AccumDir> -meanfn </means> -varfn </variances> -moddeffn <.mdef>

$Options$ :

-outmllrfn : spécifie la matrice de transformation (sortie)

-accumdir : répertoire pour accumuler les moyennes

-meanfn : spécifie les « means » du modèle HMM

-varfn : spécifie les « variances » du modèle HMM

-moddeffn : description du modèle (.mdef).

– **mllr_transform**
$Objet$ : transforme la moyenne d'un modèle HMM après la matrice de transformation obtenue par mllr_solve.
$Utilisation$ : mllr_transform -moddeffn <.mdef> -inmeanfn <1/means> -outmeanfn <2/means> -mllrmat <adapt.matrix>
$Options$ :
-moddeffn : description du modèle (.mdef)
-inmeanfn : spécifie les « means » du modèle HMM
-outmeanfn : spécifie les « means » du nouveau modèle HMM (sortie)
-mllrmat : spécifie la matrice de transformation.

## C.4 Décodage et évaluation de performances

– **TranscribeSpeechCorpus.pl**
$Objet$ : processus de décodage de la parole sur un corpus de fichiers audio. Les étapes de traitement sont spécifiées dans un fichier de configuration.
$Utilisation$ : TranscribeSpeechCorpus.pl -configFn <file> -CorpusFn <file> -resultDir <dir>
$Options$ :
-configFn <file> : spécifie le fichier de configuration, qui décrit les étapes de traitement à faire, ainsi que toute autre donnée nécessaire (modèles, lexique, ...)
-CorpusFn <file> : spécifie le corpus de fichiers audio à traiter
-resultDir <dir> : spécifie le répertoire pour les résultats (sortie).

– **TranscribeSpeechFile.pl**
$Utilisation$ : TranscribeSpeechFile.pl -config <file> -define $define -signal <wavFile> -segments <file> -res <dir.$infos.res>
$Options$ :
-config <file> : spécifie le fichier de configuration, qui décrit les étapes de traitement à faire, ainsi que toute autre donnée nécessaire (modèles, lexique, ...)
-define : permet d'ajuster des paramètres de la configuration choisie pour le décodage
-signal <wavFile> : spécifie le fichier audio a traiter
-segments <file> : spécifie la liste des segments a traiter
-res <dir.$infos.res> : spécifie le répertoire pour les résultats (sortie).

**Exemple d'utilisation :**
set Adapt = MLLR40cMAPmean # valeurs : Original, MLLR1c, MLLR1cMAP, MLLR1cMAPmean, MLLR40c, MLLR40cMAP, MLLR40cMAPmean, MAP, MAPmean
set Vocab = 0.0k # valeurs : 0.0k, 0.1k, 0.5k, 1k, 2k, 5k
set LMw = 08 # valeurs : 08, 07, 06, 05
set FillProb = 0.05 # valeurs : 0.05, 0.01, 0.005, 0.001
set define = "Adapt=${Adapt},Vocab=${Vocab},LMw=${LMw},FillProb=${FillProb}"
set infos = "Adapt_${Adapt}_Voc_${Vocab}_LMw_${LMw}_FillProb_${FillProb}"

TranscribeSpeechFile.pl \
-config ~/testDecode/reco_segments_sphinx3_v02–1mdl_16kHz_options.cfg \
-define $define \
-signal ~/Files_domotics/AudioFiles/continuous/FDJM.wav \
-segments ~/testDecode/FDJM_segments.dat \
-res /home/testDecode/FDJM_segments.$infos.res

– **AnalyzeResults.pl**
$Objet$ : analyse le résultat de la reconnaissance de la parole.
$Utilisation$ : AnalyzeResults.pl -corpusFn <file> -stmDir <dir> -ctmDir <dir> -corrCtmFn <file> -normOpt <options> -resName <fileRes>
$Options$ :
-corpusFn <file> : spécifie la liste des $audioId$ qui indique les resultats à analyser.
-stmDir <dir> : spécifie le répertoire qui contient les fichiers $.stm$
-ctmDir <dir> : spécifie le répertoire qui contient les fichiers $.ctm$ (le paramètre$-resultDir$ qui a été spécifié dans le processus $TranscribeSpeechCorpus.pl$)
-corrCtmFn <file> : correspondances à appliquer sur les fichiers $.ctm$ avant l'étape de normalisation
-normOpt <options> : spécifie les options pour la normalisation des données $.stm$ et $.ctm$ avant de leur comparaison.
-resName <fileRes> : spécifie le nom de base pour les résultats.

# Bibliographie


[1] H. CHRISTENSEN, J. BARKER, N. MA, AND P. GREEN, *The chime corpus : a resource and a challenge for computational hearing in multisource environments*, INTERSPEECH-2010, pp. 1918–1921.

[2] S.-J. DOH AND R. M. STERN, *Inter-class mllr for speaker adaptation*, in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000, pp. 1755–1758.

[3] A. FLEURY, M. VACHER, F. PORTET, P. CHAHUARA, AND N. NOURY, *A Multimodal Corpus Recorded in a Health Smart Home*, in Proceedings of the Workshop on Multimodal Corpora : Advances in Capturing, Coding and Analyzing Multimodality in conjunction with LREC 2010, Valetta, Malta, May 2010, pp. 99–105.

[4] S. GALLIANO, G. GRAVIER, AND L. CHAUBARD, *The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts*, INTERSPEECH-2009, pp. 2583–2586.

[5] J.-P. HATON, C. CERISARA, D. FOHR, Y. LAPRIE, AND K. SMAILI, *Reconnaissance Automatique de la Parole. Du signal a son interpretation*, UniverSciences (Paris) - ISSN 1635-625X, DUNOD, 2006.

[6] R. JUSTO, O. SAZ, V. GUIJARRUBIA, A. MIGUEL, M. I. TORRES, AND E. LEIDA, *Improving dialogue systems in a home automation environment*, in Proceedings of the 1st international conference on Ambient media and systems, Ambi-Sys '08, ICST, Brussels, Belgium, Belgium, 2008, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp. 2 :1–2 :6.

[7] B. KOTNIK AND Z. KACIC, *A noise robust feature extraction algorithm using joint wavelet packet subband decomposition and ar modeling of speech signals*, Signal Process., 87 (2007), pp. 1202–1223.

[8] F. LAURI, I. ILLINA, AND D. FOHR, *Adaptation mllr pour des hmms*, in Quatriemes Rencontres Jeunes Chercheurs en Parole - RJC 2001, Mons, Belgique, 2001, pp. 90–93. Colloque avec actes sans comite de lecture nationale.

[9] C. J. LEGGETTER AND P. C. WOODLAND, *Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models*, Computer Speech and Language, 9 (1995), pp. 171–185.

[10] D. MACHO, L. MAUUARY, B. NOE, Y. M. CHENG, D. EALEY, D. JOUVET, H. KELLEHER, D. PEARCE, AND F. SAADOUN, *Evaluation of a noise-robust dsr front-end on aurora databases*, in INTERSPEECH, 2002.

[11] S. MOLLER, F. GODDE, AND M. WOLTERS, *Corpus analysis of spoken smart-home interactions with older users*, in Proceedings of the Sixth International Language Resources and

Evaluation (LREC 08), B. M. J. M. J. O. S. P. D. T. Nicoletta Calzolari (Conference Chair), Khalid Choukri, ed., Marrakech, Morocco, may 2008, European Language Resources Association (ELRA). http ://www.lrec-conf.org/proceedings/lrec2008/.

[12] J. PICONE, *Continuous speech recognition using hidden markov models*, IEEE ASSP Magazine, 7 (1990), pp. 26–41.

[13] J. PICONE, *Signal modeling techniques in speech recognition*, vol. 81, September 1993, pp. 1215–1247.

[14] J. PINQUIER, J.-L. ROUAS, J. MAUCLAIR, AND R. ANDRE-OBRECHT, *Detection de la parole et de la musique : fusion de deux approches*, in 19e Colloque GRETSI sur le traitement du signal et des images (GRETSI 2003), vol. 3, Paris, France, September 2003, Telecom-Paris, pp. 78–81.

[15] L. R. RABINER, *A tutorial on hidden markov models and selected applications in speech recognition*, in IEEE Proceedings, 1989, pp. 257–286.

[16] C. RAYMOND, *Decodage conceptuel : co-articulation des processus de transcription et comprehension dans les systemes de dialogue*, PhD thesis, Universite d'Avignon et des Pays de Vaucluse, 2005.

[17] H. SAKOE AND S. CHIBA, *Dynamic programming algorithm optimization for spoken word recognition*, IEEE Transactions on Acoustics, Speech, and Signal Processing, (1978), pp. 43–49.

[18] A. SORIN, T. RAMABADRAN, D. CHAZAN, R. HOORY, M. MCLAUGHLIN, D. PEARCE, F. C. WANG, AND Y. ZHANG, *The etsi extended distributed speech recognition (dsr) standards : client side processing and tonal language recognition evaluation*, in International Conference on Acoustics, Speech, and Signal Processing, 2004.

[19] THIEBAUT-BRODIER, *Domotique : securite - confort - economies*, Publitronic Elektor, 2005.

[20] M. VACHER, A. FLEURY, F. PORTET, J.-F. SERIGNAT, AND N. NOURY, *Complete Sound and Speech Recognition System for Health Smart Homes : Application to the Recognition of Activities of Daily Living*, Intech Book, February 2010, pp. 645 – 673.

[21] J.-C. WANG, H. P. LEE, J.-F. WANG, AND C.-B. LIN, *Robust environmental sound recognition for home automation*, IEEE T. Automation Science and Engineering, 5 (2008), pp. 25–31.

[22] P. C. WOODLAND, *Speaker adaptation for continuous density hmms : A review*, pp. 11–19.

[23] D. WU, M. TANAKA, R. CHEN, L. OLORENSHAW, M. AMADOR, AND X. MENENDEZ-PIDAL, *A robust speech detection algorithm for speech activated hands-free applications*, Acoustics, Speech, and Signal Processing, IEEE International Conference on, 4 (1999), pp. 2407–2410.