

Reconnaissance de parole avec prise de son distante pour la domotique

Encadrants:

Dominique Fohr

Irina Illina

Denis Jouvét

Equipe

PAROLE

Etudiante:

Luiza Orosanu

Nancy, le 26 Juin 2011

Plan

- Contexte et problématique
- Notions générales
- Contexte du travail
- Étude paramétrage et modélisation
- Corpus parole
- Résultats d'expérimentations
- Conclusion

Plan



Contexte et problématique

La reconnaissance de la parole [2006; J.-P. HATON, C. CERISARA, D. FOHR, Y. LAPRIE, AND K. SMAILI]

Objectifs:

- la transcription
- la traduction
- faciliter l'exécution de commandes

Contraintes:

- la qualité du signal capté par les microphones
- la performance des outils de reconnaissance

Contexte et problématique

Domaines d'application:

- un environnement téléphonique avec une prise de son effectuée par le combiné téléphonique
- un environnement calme avec une prise de son de bonne qualité (telle que micro-casque) par exemple pour la dictée vocale
- un environnement domestique [2005; THIEBAUT-BRODIER] avec une prise de son à distance




La proximité au microphone est une contrainte nécessaire au bon fonctionnement du système de reconnaissance de la parole

Contexte et problématique

Difficultés liées à l'application de la reconnaissance de la parole au domaine de la domotique :

- variabilité de la parole
- éloignement des locuteurs aux microphones
- bruit ambiant
- faire la distinction entre les commandes adressées à la centrale domotique et les conversations (résidents discutant entre eux, ...)
- atteindre un niveau de performance acceptable pour l'utilisateur

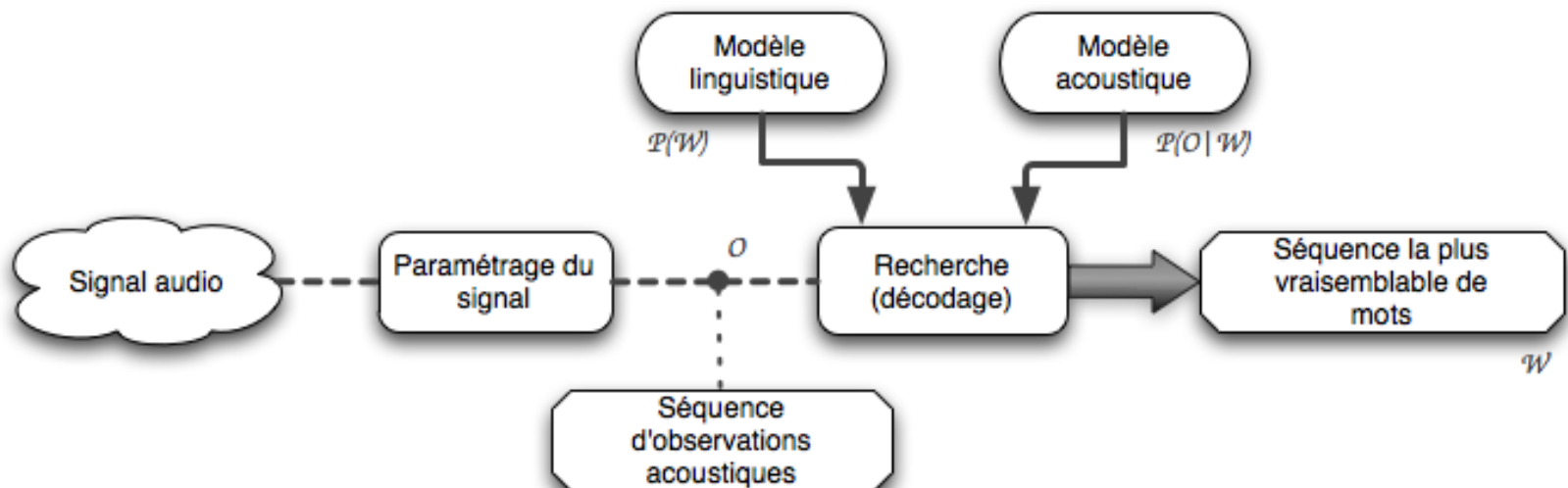
Plan



**Notions
générales**

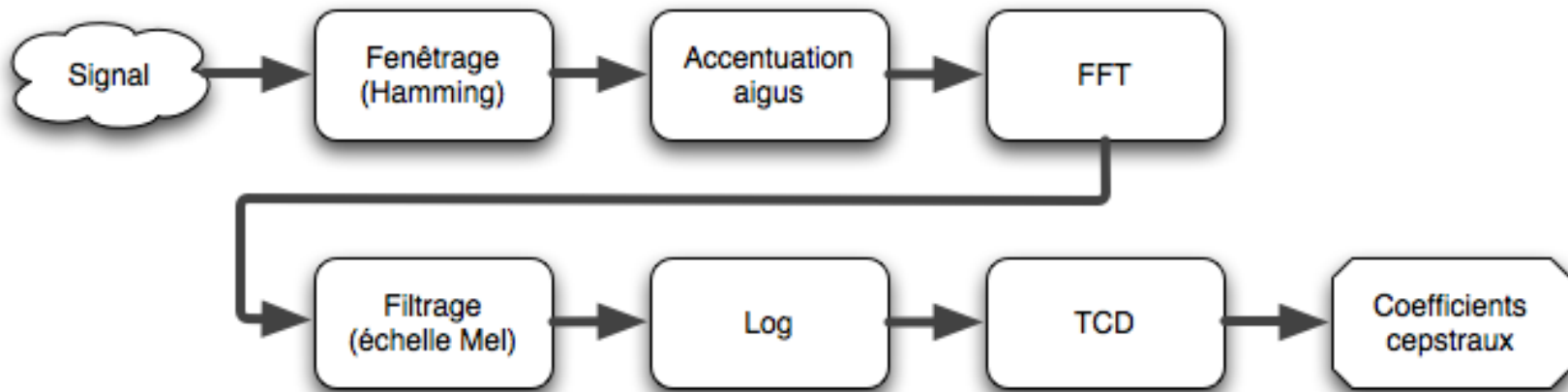
Modélisation pour la reconnaissance

Architecture d'un système de reconnaissance vocale:



Analyse du signal vocal

Paramétrage du signal audio (Analyse cepstrale MFCC)



Analyse Aurora : débruitage + MFCC

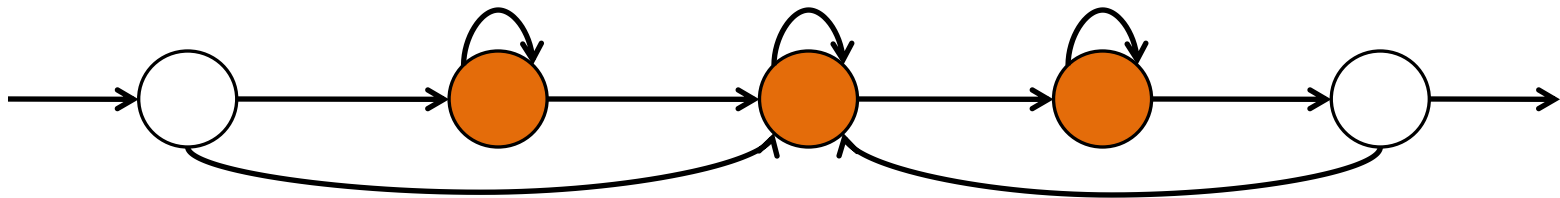
Modèles de Markov caché (HMM)

Modèles de Markov:

- automates probabilistes à états finis
- définis par deux processus stochastiques :
 - transitions
 - émission d'observations

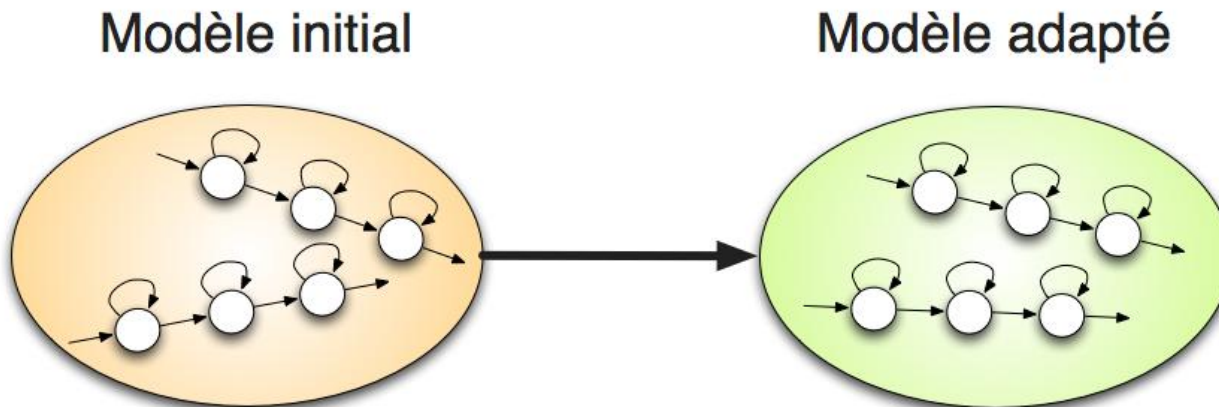
Application au traitement de la parole:

- un modèle de Markov associé à chaque unité de parole
- l'émission d'une observation dépend seulement de l'état courant



Adaptation de modèles HMM

- Mise à jour de paramètres du modèle acoustique (HMM) afin d'améliorer la modélisation pour un nouveau locuteur ou pour un nouvel environnement
- Méthodes couramment utilisées [2001, Woodland]:
 - maximum a posteriori (MAP)
 - régression linéaire (MLLR)



Évaluation

Évaluation de performance d'un système de reconnaissance:

- Comparaison entre:
 - transcription de référence: les mots qui ont été prononcés
 - transcription hypothèse: les mots qui ont été reconnus par le système
- Critère d'évaluation par taux d'erreur:

$$WER = \frac{Substitutions + Omissions + Insertions}{Nombre\ de\ mots\ référence}$$

Plan



Contexte du travail

- **Objectif :**
 - évaluer les performances d'un système de reconnaissance avec prise de son distante
- **Descriptif du travail:**
 - enregistrement de données à distance
 - synchronisation des transcriptions de référence
 - déterminer la configuration conduisant à une performance acceptable pour l'utilisateur

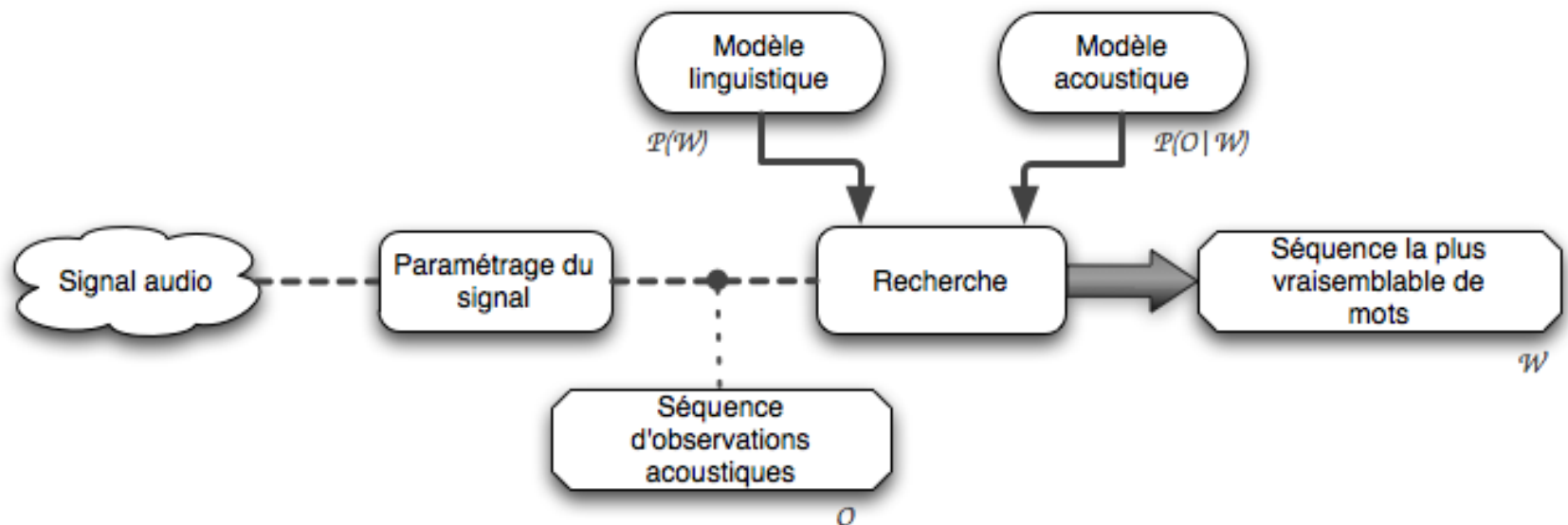
Plan

A 3D rendered orange figure stands on a light brown surface, holding a rectangular sign with both hands. The sign has a light beige background and a thin brown border. The text on the sign is in a bold, orange, sans-serif font.

**Étude paramétrage
et modélisation**

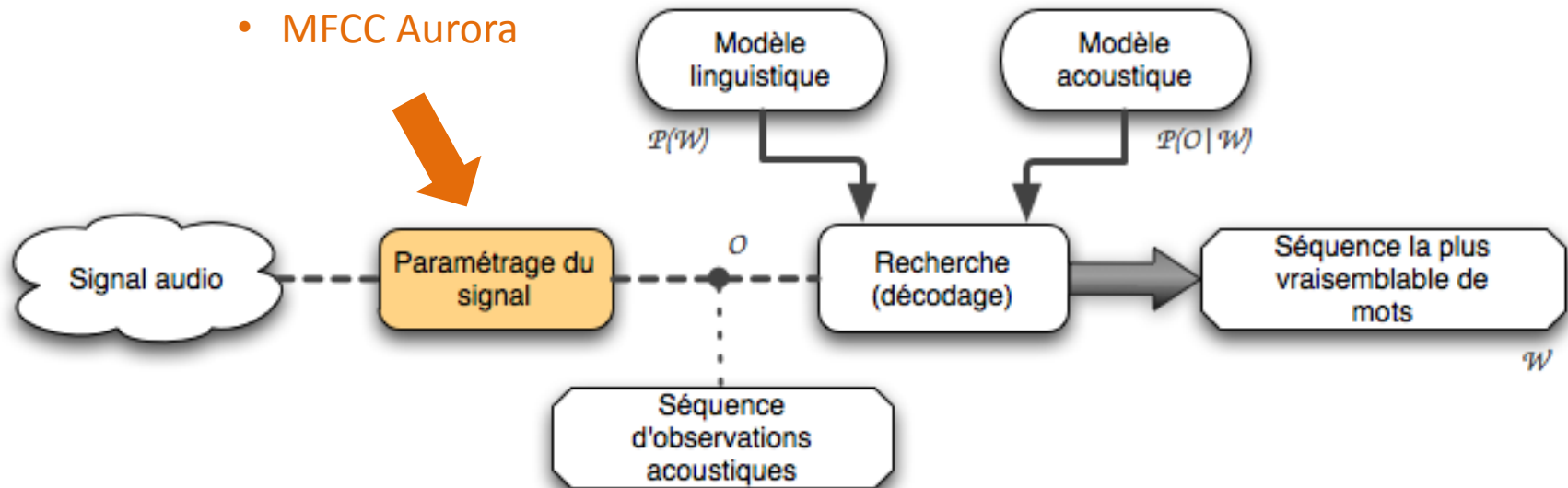
Étude paramétrage et modélisation

Analyse de l'influence des paramètres du systèmes de reconnaissance:



Analyse acoustique

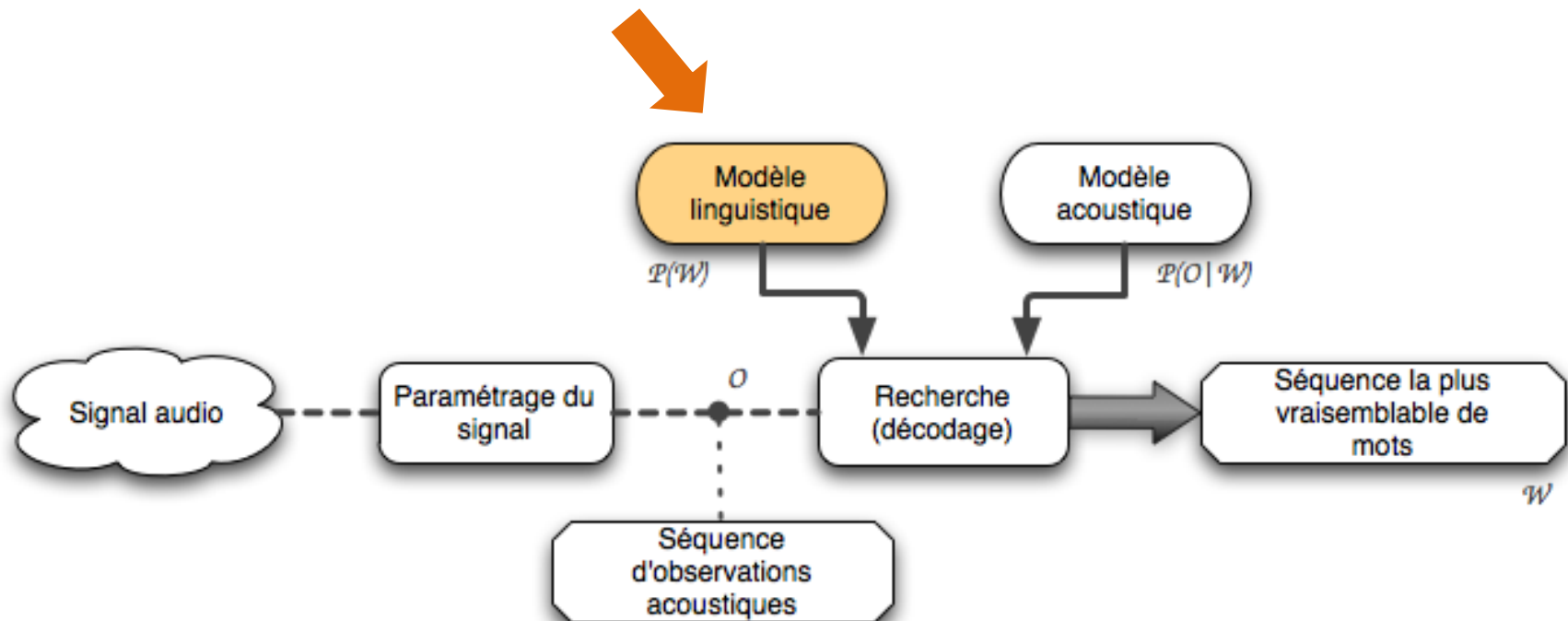
- MFCC Sphinx
- MFCC Aurora



- impact de l'analyse acoustique du signal audio

Vocabulaire

- $\text{voc} = \text{domotique} + 0.0\text{k}$
- $\text{voc} = \text{domotique} + 0.1\text{K}$
- $\text{voc} = \text{domotique} + 0.5\text{k}$
- $\text{voc} = \text{domotique} + 1\text{k}$

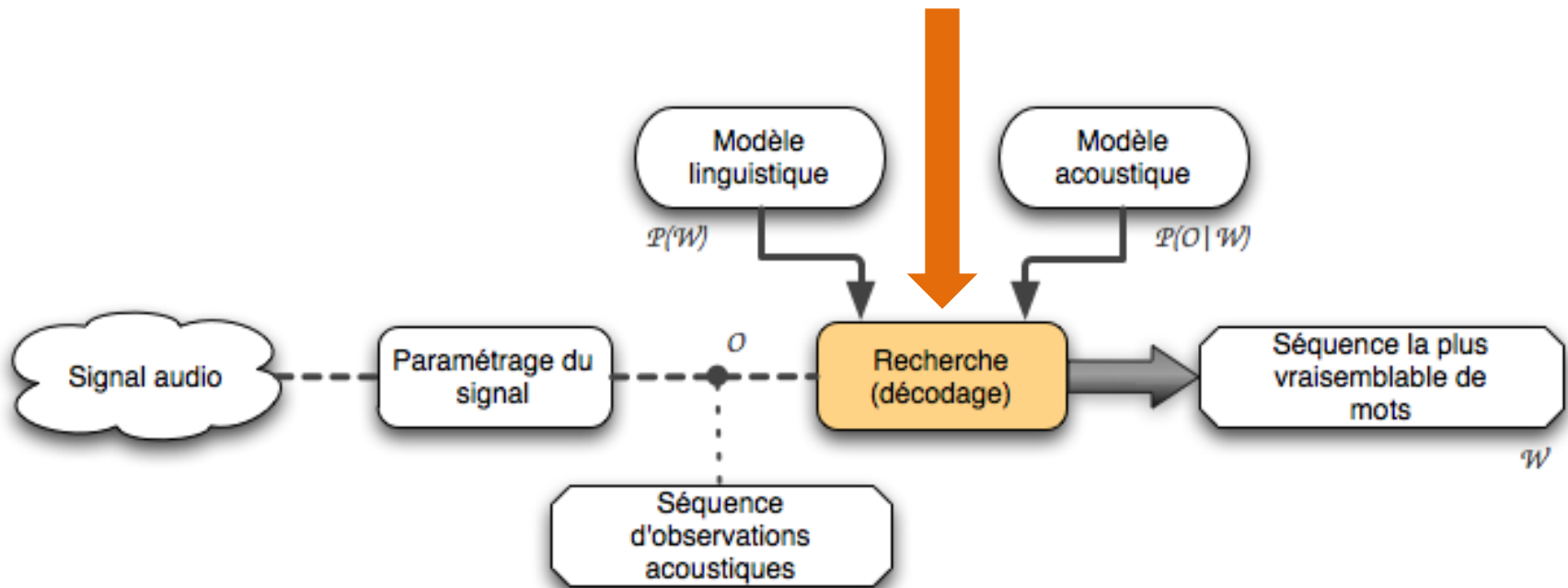


- **impact du vocabulaire**

Poids du modèle de langage

$$\hat{W} \equiv \underset{W}{\operatorname{ArgMax}} P(O|W)P(W)^{l_w}$$

- $l_w = 08$
- $l_w = 06$
- $l_w = 07$
- $l_w = 05$

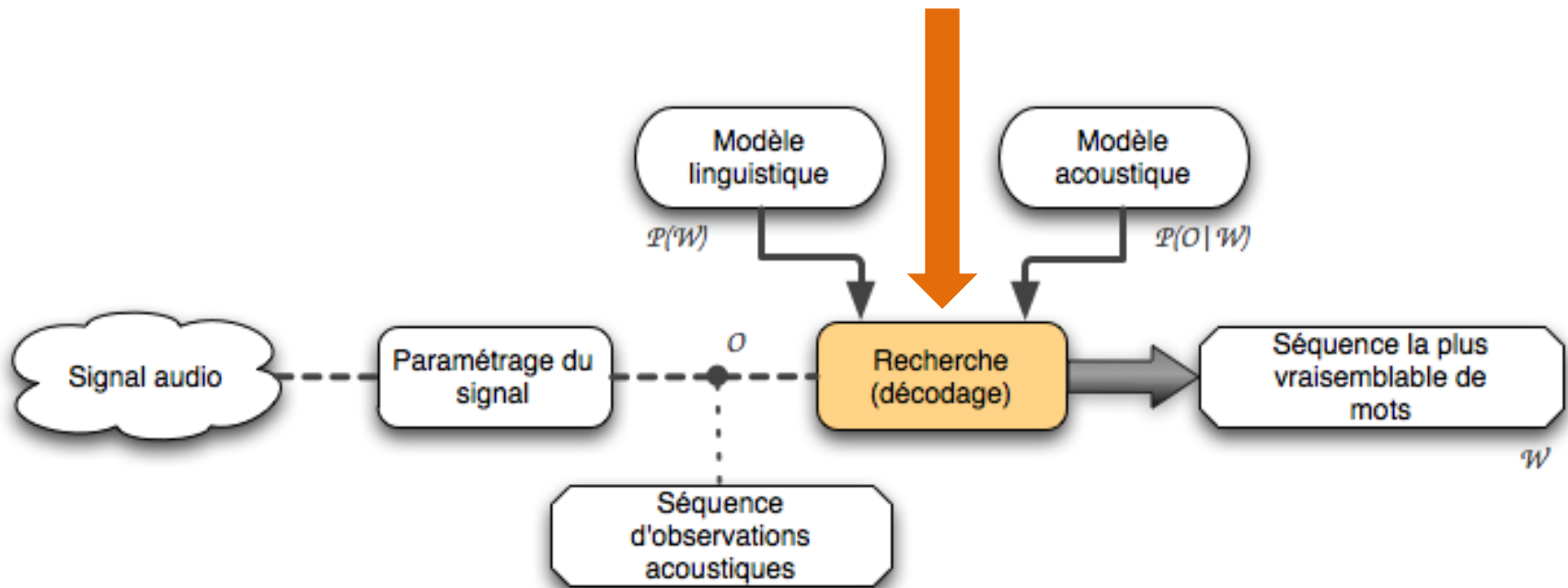


- **impact de poids du modèle de langage**

Probabilité des fillers

fillers = modèles de bruit

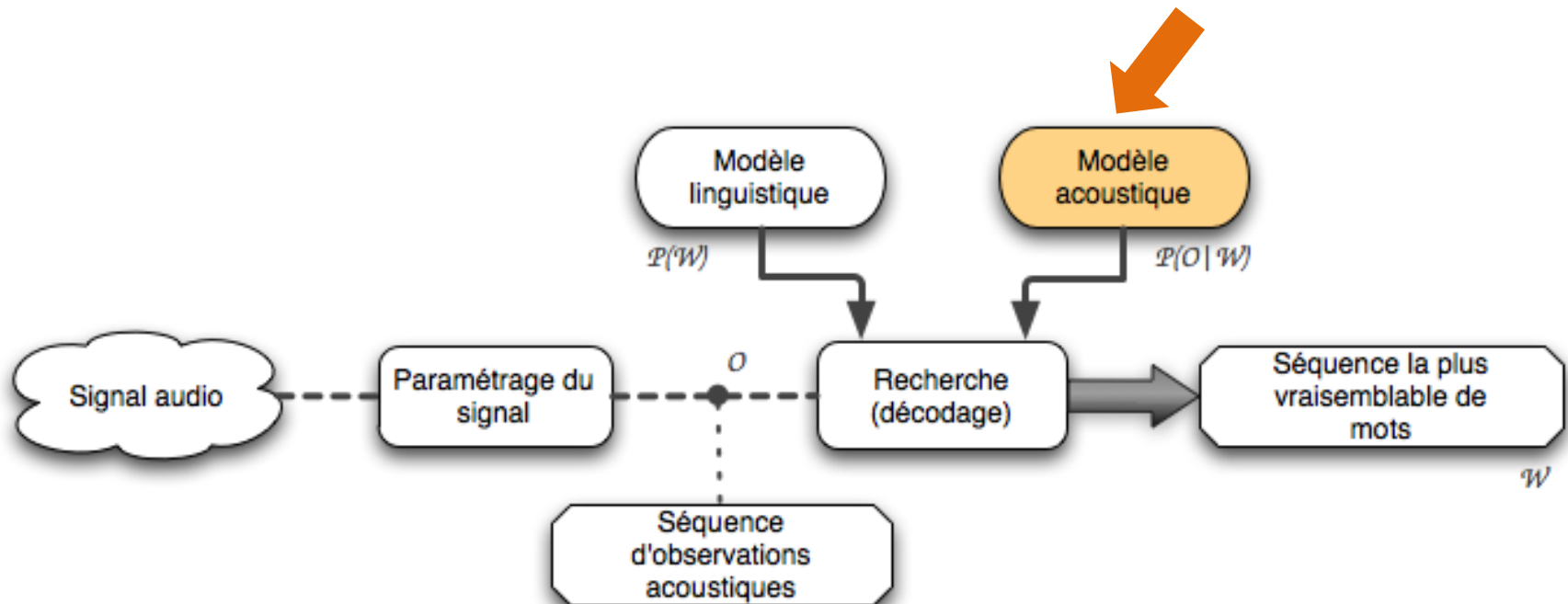
- FillProba = 0.05
- FillProba = 0.01
- FillProba = 0.005
- FillProba = 0.001



- **impact de la probabilité des fillers**

Adaptation modèles acoustiques

- Original
- adaptation MAP
- adaptation MLLR
- combinaisons MLLR + MAP



- impact de l'adaptation de modèles acoustiques

Plan



Corpus domotique - spécification & enregistrement

Informations :

- correspondant à une liste de commandes domotiques (avec et sans le mot clé)
- Exemple de commande: « *Majordome, allume la lumière.* »

Conditions d'enregistrement :

- environnement: sans réverbération, peu de bruit ambiant (SNR de $\sim 15\text{dB}$)
 - les « fichiers en continu »: à une distance de 1 mètre
 - les « fichiers segmentés »: à une distance de 40cm

Utilisation :

- les « fichiers en continu »: données de test
- les « fichiers segmentés »: données de référence pour la synchronisation de transcriptions

Corpus domotique - transcriptions de référence

On dispose de: transcriptions de référence pour les « fichiers segmentés »

On a besoin de: transcriptions de référence pour les « fichiers en continu »

Synchronisation de transcriptions de référence:

- alignement élastique entre les « fichiers segmentés » et les « fichiers en continu »
 - trouver la position de chaque phrase dans le signal enregistré en continu à distance
 - sélectionner uniquement les phrases qui ont été bien prononcées



Corpus ESTER2 (Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques)

Informations :

- contient des bulletins d'information, manuellement transcrits
- divisé dans un ensemble d'apprentissage et un autre de développement

Conditions d'enregistrement :

- environnement: réel
- à une distance de 1 mètre

Utilisation :

- données d'adaptation

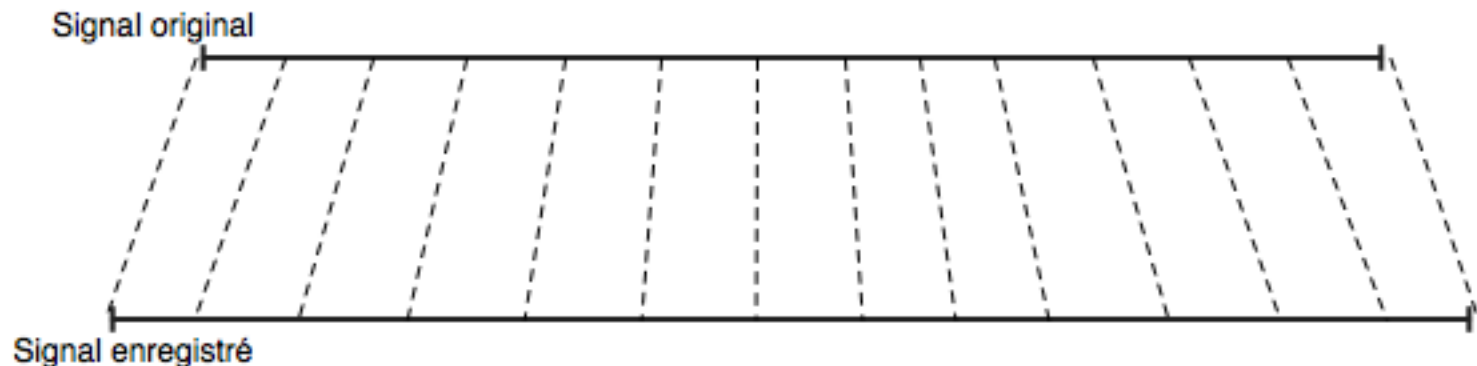
Corpus ESTER2 - transcriptions de référence

On dispose de: transcriptions de référence pour les fichiers audio d'ESTER2

On a besoin de: transcriptions de référence pour les fichiers enregistrés à distance

Synchronisation de transcriptions de référence:

- alignement élastique entre les fichiers originaux et ceux enregistrés à distance
 - trouver la position de chaque segment de parole dans le signal enregistré à distance



Corpus CHIME (Computational Hearing in Multisource Environments)

Informations :

- contient des commandes artificielles
- divisé dans un ensemble d'apprentissage et un autre de développement

Conditions d'enregistrement :

- environnement: réel, réverbérant, plus ou moins bruité (valeurs SNR: -6dB, -3dB, 0dB, 3dB, 6dB, 9dB)
 - à distance

Utilisation :

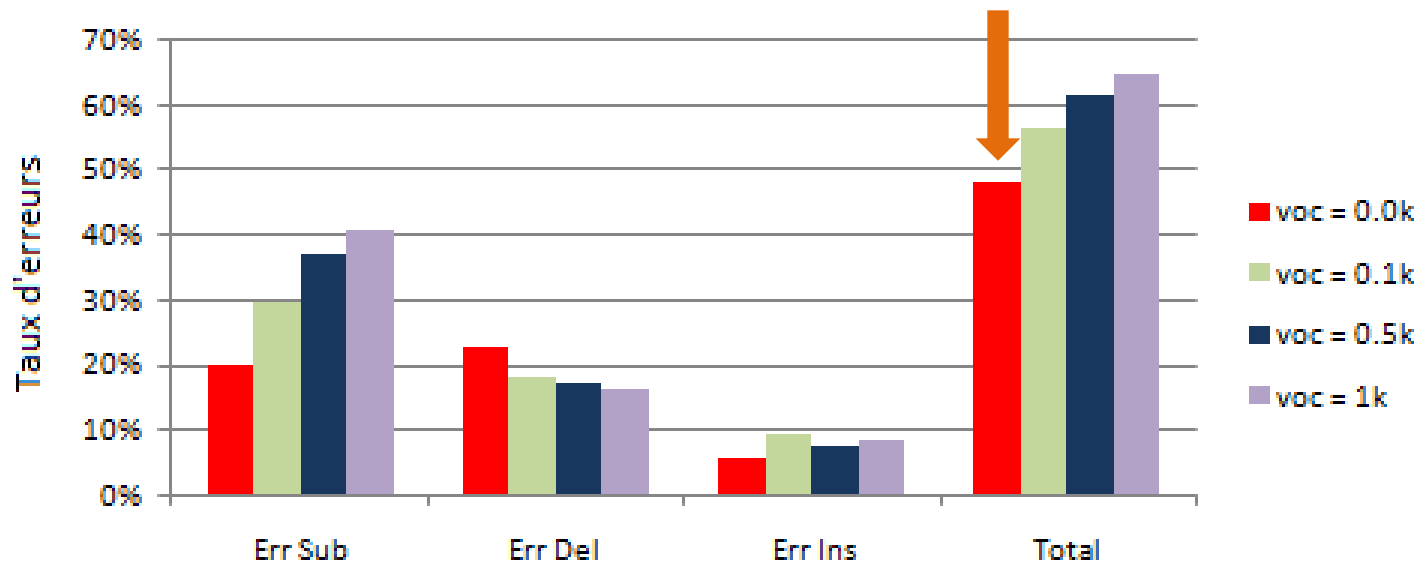
- données de test bruitées

A 3D rendered orange figure stands on a light brown surface, holding a rectangular sign with a thin orange border. The figure is positioned behind the sign, with its hands visible at the top and sides. The background is a plain, light yellowish-white.

**Résultats
d'expérimentations**

Vocabulaire

- Configuration**
- analyse acoustique MFCC Sphinx
 - le modèle acoustique non-adapté d'ESTER2
 - différentes variantes de vocabulaire

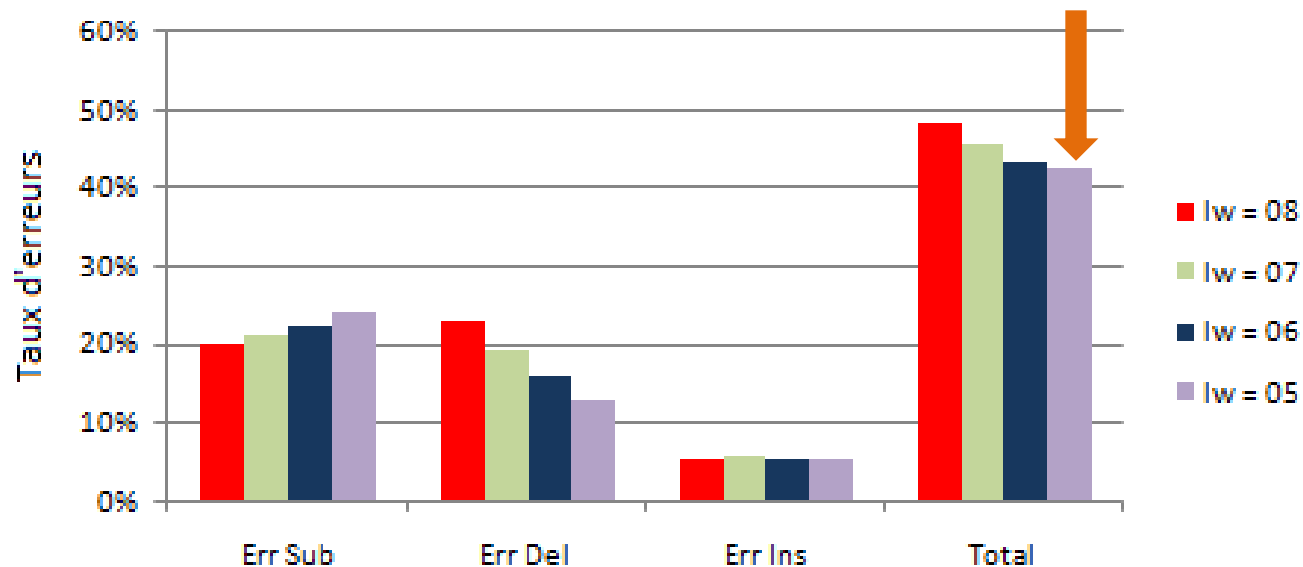


➡ meilleurs résultats (sur le corpus domotique) obtenus avec le vocabulaire limité à celui des commandes domotiques

Poids du modèle de langage

Configuration

- analyse acoustique MFCC Sphinx
- le modèle acoustique non-adapté d'ESTER2
- le vocabulaire domotique
- différentes variantes de poids du modèle de langage

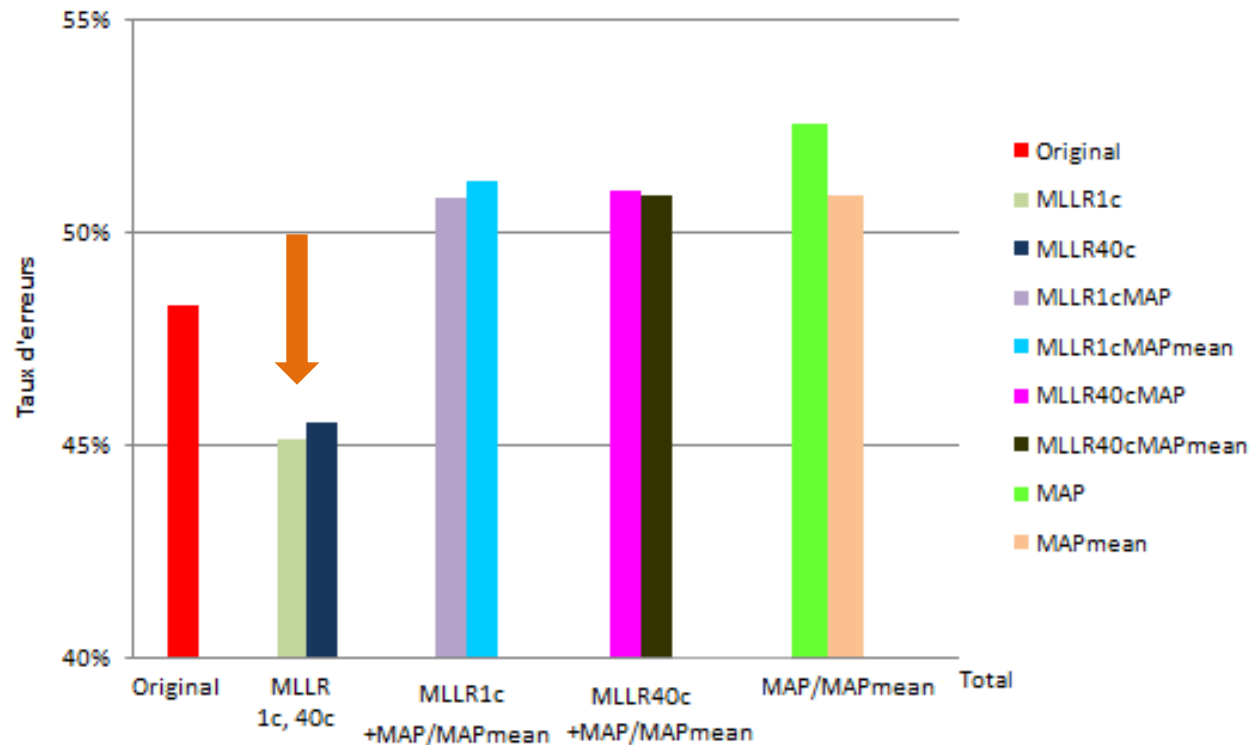


➡ la diminution du *poids du modèle de langage* améliore les résultats

Adaptation de modèle acoustique

Configuration

- analyse acoustique Sphinx
- le vocabulaire domotique
- différentes variantes de modèles acoustiques adaptés



➡ les meilleurs résultats obtenus (sur le corpus domotique) avec les modèles adaptés par régression linéaire avec 1 ou 40 classes

Autres paramètres

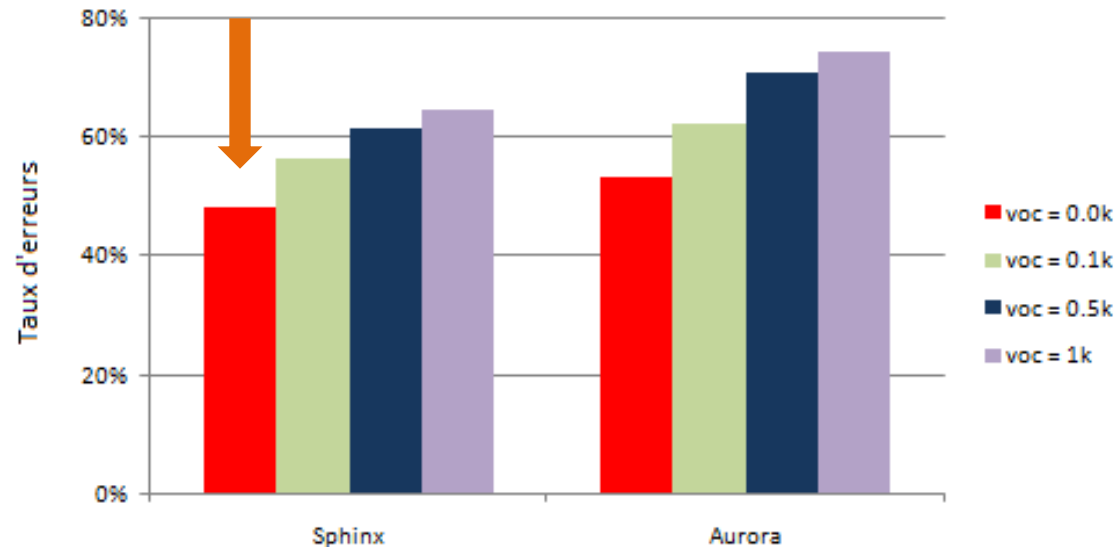
Résultats avec des autres tests:

- la variation de *probabilité des fillers* n'apporte aucun gain au résultats
- l'utilisation d'un mot clé long et sonore au début de chaque phrase donne une petite amélioration des résultats
- les erreurs de reconnaissance sur le mot clé sont faibles (de l'ordre 10%)

Analyse acoustique

Configuration

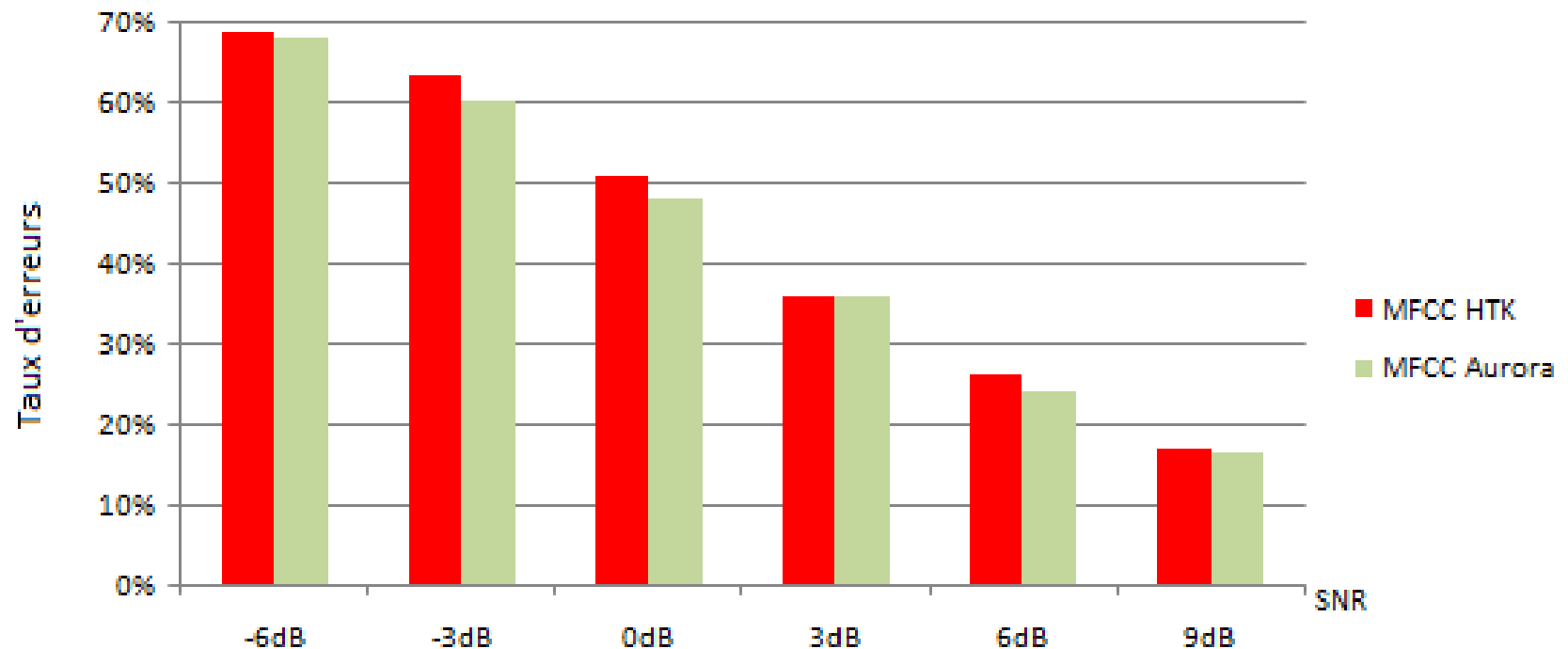
- analyse acoustique MFCC Sphinx et MFCC Aurora (+phase debruitage)
- le modèle acoustique non-adapté d'ESTER2
- différentes variantes de vocabulaire



➡ le paramétrage MFCC Aurora n'améliore pas les résultats obtenus par rapport au paramétrage MFCC Sphinx

Analyse acoustique

- le paramétrage MFCC Aurora sur le corpus CHIME



➔ le paramétrage Aurora améliore les résultats pour le corpus CHIME

Plan



Rappel des questions posées

Questions posées au début du stage:

- 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 ?

Conclusion

D'après les résultats de nos expérimentations, afin d'améliorer la performance de reconnaissance il faut :

- utiliser le vocabulaire limité à la domotique,
- diminuer le poids du modèle de langage
- 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.

Perspectives

Quelques points qu'il reste encore à aborder:

- le problème de détection des zones parole / non parole en milieu bruyant
- l'amélioration du modèle de langage, en le fabricant à partir d'un grand nombre d'exemples de commandes domotiques.

A 3D rendered orange figure, resembling a stylized person, stands and holds a large, rectangular, light-colored sign with a thin orange border. The figure is positioned behind the sign, with its hands visible at the top and sides. The background is a plain, light yellow gradient.

**Merci pour
votre attention!**

A 3D rendered orange figure stands on a light-colored surface, holding a large, rectangular, light-colored sign with a thin orange border. The figure is positioned behind the sign, with its hands visible at the top and sides. The sign has the word "Questions?" written on it in a bold, orange, sans-serif font. The background is a plain, light yellow gradient.

Questions?