L. Orosanu, D. Jouvet, D. Fohr,
I. Illina, A. Bonneau

Équipe PAROLE, INRIA – LORIA
615 Rue du Jardin Botanique
54600 Villers-les-Nancy

Detection de transcriptions incorrectes de parole non-native dans le cadre de l'apprentissage de langues étrangères

6 Juin 2012

Plan

- **Contexte et problématique**
- Méthodologie
- Expériences et résultats
- Conclusions

Contexte et problématique

Apprentissage des langues étrangères

- utilisation des technologies de **reconnaissance de la parole**
- pour détecter et signaler les **erreurs ou défauts de prononciation**

Contexte et problématique

- Il arrive que le signal acoustique **ne corresponde pas** à la phrase attendue

- paroles parasites
- problème de capture du son
- …

➡ Le système doit être capable de détecter les entrées incorrectes : signal audio ne correspondant pas à la phrase attendue (le texte attendu)

Plan

- Contexte et problématique
- **Méthodologie**
  - **Exemple**
  - **Critères pour la décision**
  - **Classification transcription correcte / incorrecte**
- Expériences et résultats
- Conclusions

Méthodologie

Loria
Laboratoire lorrain de recherche en informatique et ses applications

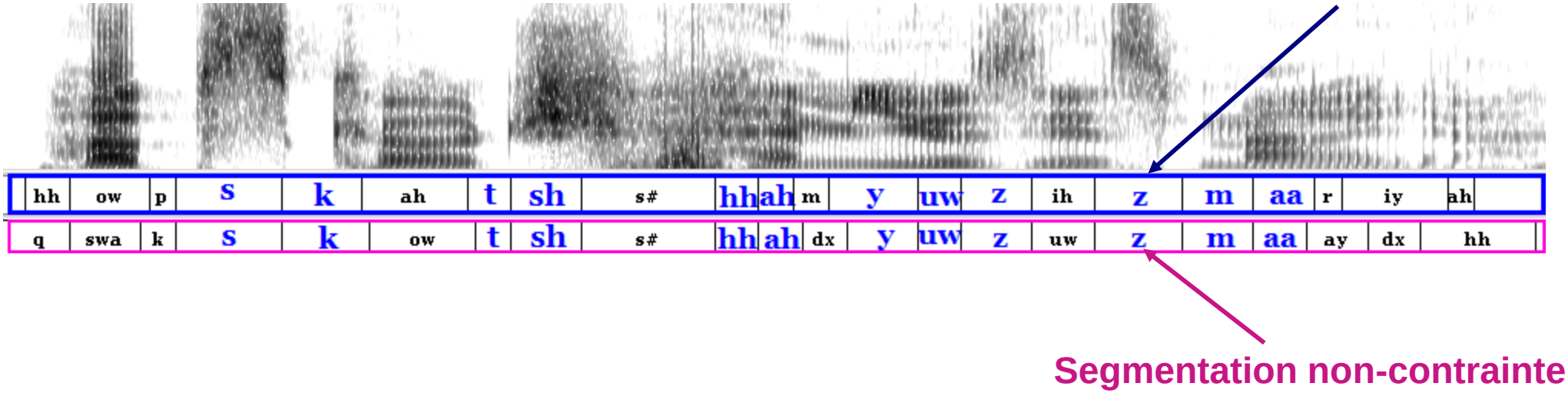
- **Objectif**
  - **rejeter** les entrées **incorrectes**
  - **accepter** les entrées **correctes**
- **Approche:** Comparaison entre
  - un **alignement contraint** par le texte attendu
  - une **segmentation non-contrainte** realisée par un décodage phonétique

Défi: tolérer les défauts de prononciation inhérents à la parole non-native

Exemple

- **Exemple d'une entrée correcte**

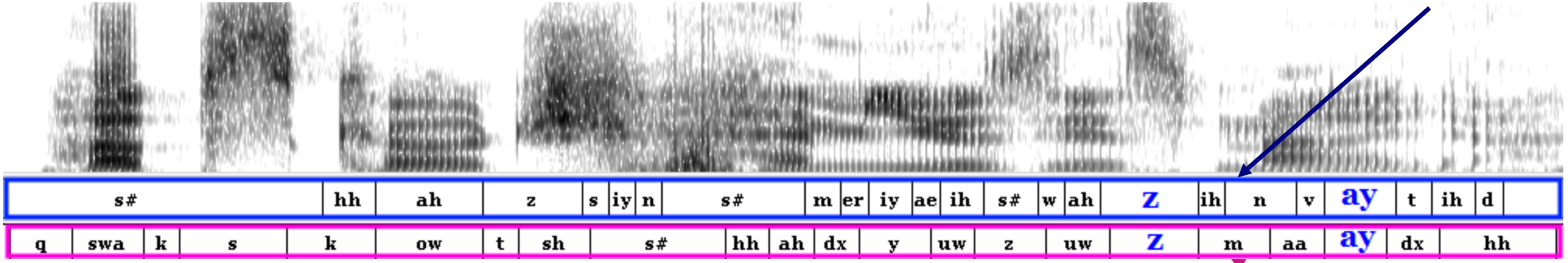
Phrase prononcée = phrase attendue ("*Hopscoch amuses Maria.*")



Exemple

- **Exemple d'une entrée incorrecte**

Phrase prononcée ≠ phrase attendue



Phrase prononcée: "*Hopscoch amuses Maria.*"

Phrase attendue: "*He has seen Maria. He was invited.*"

Critères pour la décision

- Exemple entrée correcte

hh	ow	p	s	k	ah	t	sh	s#	hh	ah	m	y	uw	z	ih	z	m	aa	r	iy	ah	
q	swa	k	s	k	ow	t	sh	s#	hh	ah	dx	y	uw	z	uw	z	m	aa	ay	dx	hh	

- Exemple entrée incorrecte

s#			hh	ah	z	s	iy	n	s#		m	er	iy	ae	ih	s#	w	ah	z	ih	n	v	ay	t	ih	d	
q	swa	k	s	k	ow	t	sh	s#		hh	ah	dx	y	uw	z	uw	z	m	aa	ay	dx	hh				

Quels critères de comparaison choisir afin de determiner si l' entrée est correcte ou non?

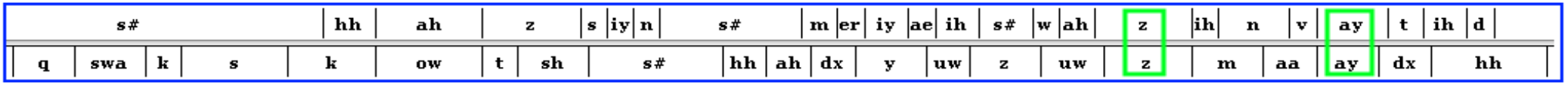
Critères pour la décision

1. Critère associé aux phonèmes

= pourcentage de segments phonétiques qui ont le même label dans les deux alignements et dont au moins une limite diffère de moins de 20 ms

Entrée correcte: 12/21 => ***57%***

Entrée incorrecte: 2/21 => ***8%***



- Les segments de non-parole sont ignorés

Critères pour la décision

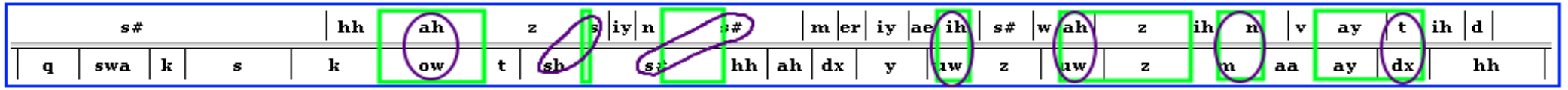
2. Critère associé aux trames

= pourcentage de trames ayant leurs étiquettes appartenant à la même classe

(6 classes phonétiques : voyelles, semi-voyelles, fricatives, affriquées, plosives, nasales)

Entrée correcte: ***85%***

Entrée incorrecte: ***52%***



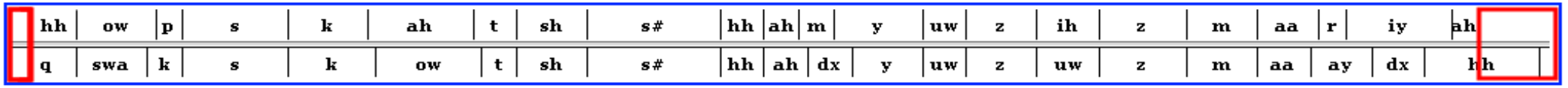
- Les segments de non-parole sont pris en compte

Critères pour la décision

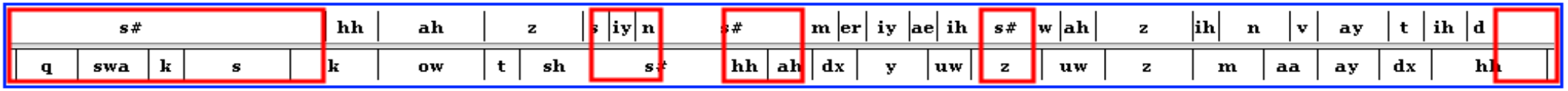
3. Critère associé aux zones de non-parole

= différence entre les recouvrements des segments de non-parole

Entrée correcte: ***4%***

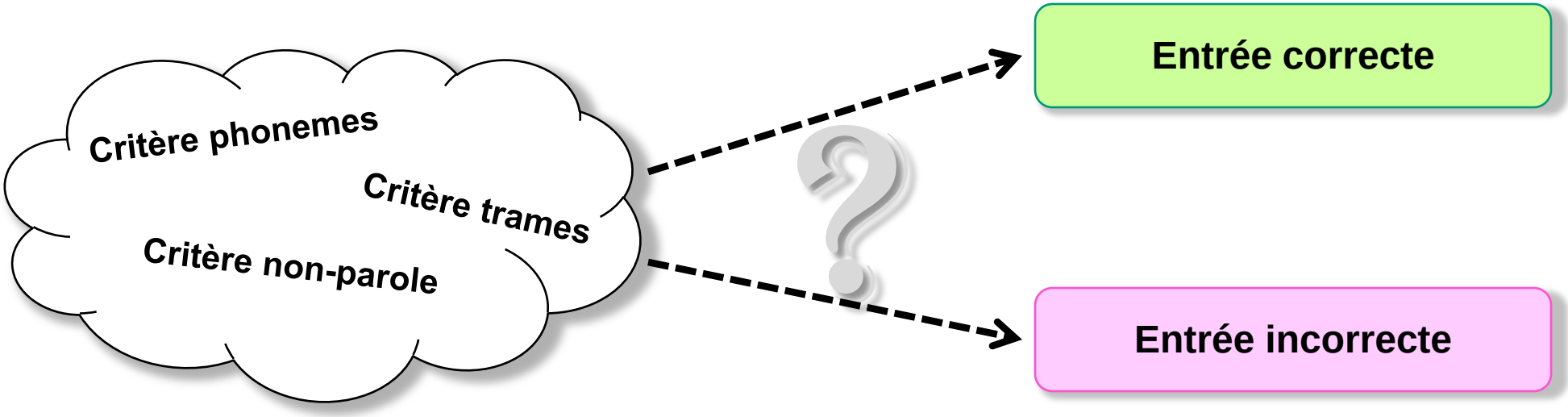


Entrée incorrecte: ***19%***



Classification

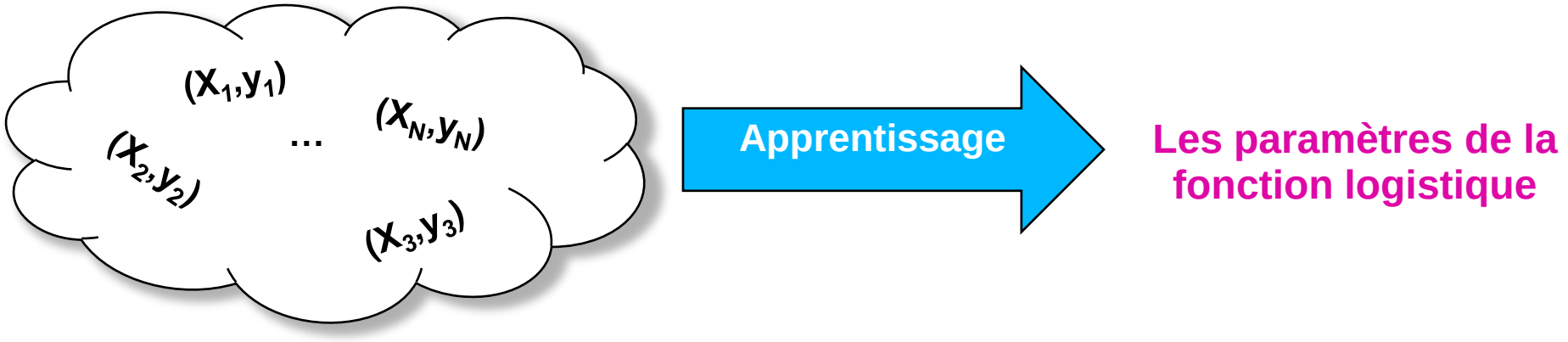
- Classification entre deux classes: **correcte** / **incorrecte**



Classification

Apprentissage

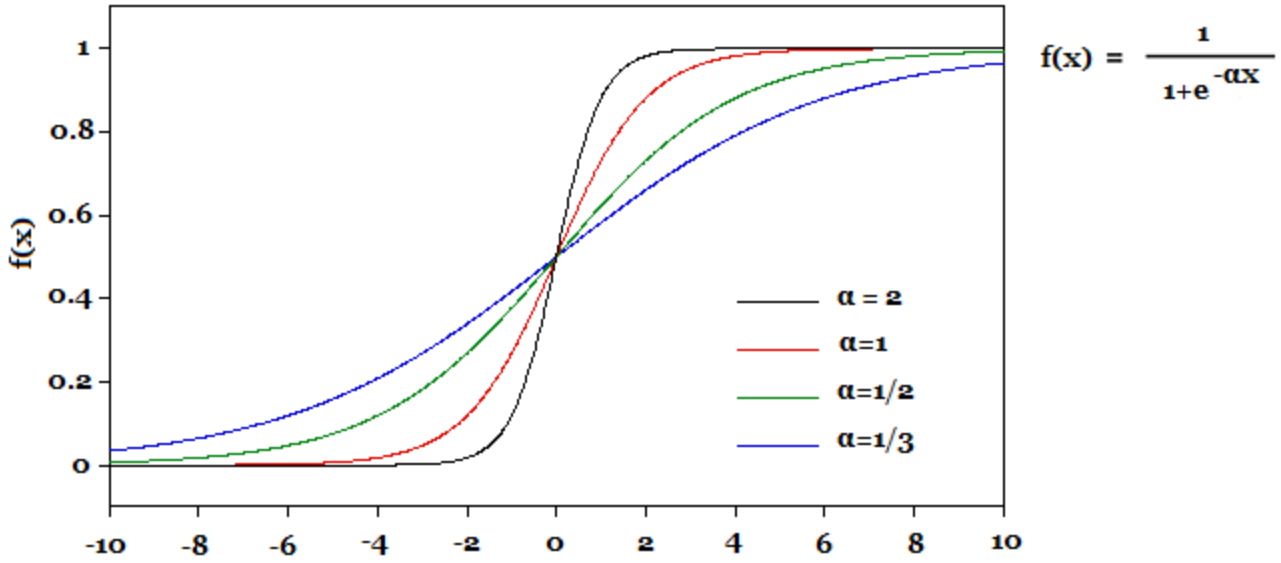
- Données: D = $\{X_i, y_i\}$, $1 \leq i \leq N$
  - $X_i = <x_1, x_2, x_3>$ les informations (critères de décision) concernant l'entrée *i* à classifier
  - y_i = 1 (entrée correcte) ou 0 (entrée incorrecte)
  - N = le nombre des éntrées (correctes & incorrectes) dans le corpus d'apprentissage



Classification

- Probabilité d'appartenance à une classe parmi deux (***modèle de la régression logistique***)

$$P(1|\overline{X},\alpha) = f(\overline{X}) = \frac{1}{1+\exp\left(-\left(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\alpha_3 x_3\right)\right)}$$

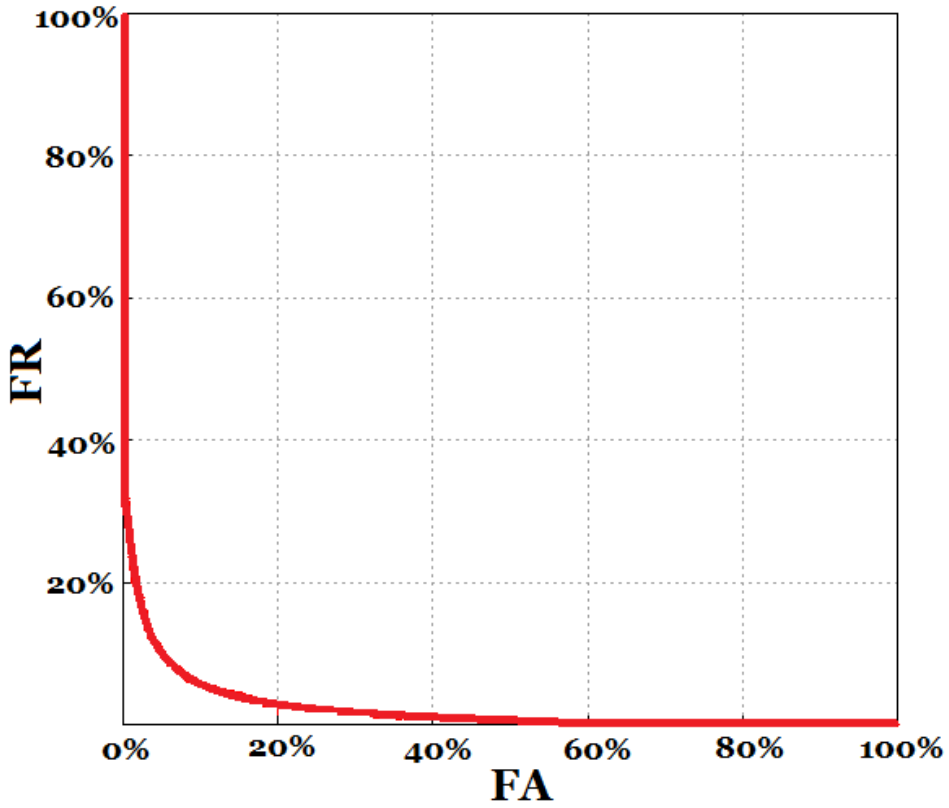


- Paramètres α estimés en minimisant la fonction d'erreur (***optimisation par descente du gradient***)

$$E = -\sum_{i=1}^{N}\left(y_i \cdot \ln\left(f(\overline{X}_i)\right)+\left(1-y_i\right)\cdot \ln\left(1-f(\overline{X}_i)\right)\right)$$

Classification

Évaluer la performance de la tâche de classification:



$$FA = \frac{\#\text{ entrées incorrectes acceptées à tort}}{\#\text{ entrées incorrectes}}$$

$$FR = \frac{\#\text{ entrées correctes rejetées à tort}}{\#\text{ entrées correctes}}$$

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{1-FA} + \frac{1}{1-FR}\right)$$

Plan

- Contexte et problématique
- Méthodologie
- **Expériences et résultats**
  - **Données**
  - **Configuration**
  - **Étude du paramétrage**
  - **Résultats**
- Conclusions

Données

- Expériences menées sur données natives et non-natives (Projet *INTONALE*)
- Corpus natif
  - ~1500 énoncés anglais
  - 22 locuteurs anglais (15 femmes, 7 hommes)
- Corpus non-natif
  - ~800 énoncés anglais
  - 34 locuteurs français (29 femmes, 5 hommes)
- Une moitié de données pour l'apprentissage (des fonctions logistiques) et l'autre pour l'évaluation

Configuration

- Pour le décodage des signaux audio: *HTK*
- Analyse accoustique

  MFCC : 12 coefficients *MFCC* + le logarithme de l'énergie par trame
- Modèles accoustiques

  HMM : chaque état modélisé par un mélange de 16 gaussiennes; appris sur *TIMIT*
- Deux lexiques

  Natif : inclut seulement les variantes natives de prononciation (*CMU*)

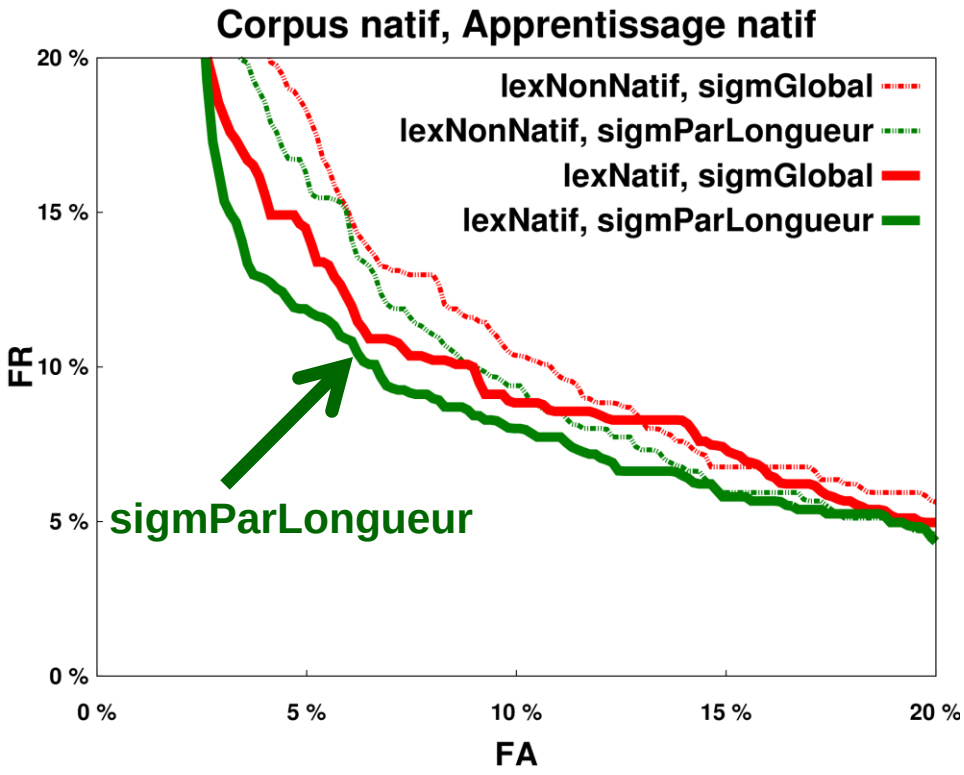
  Non-natif : inclut en plus des variantes non-natives

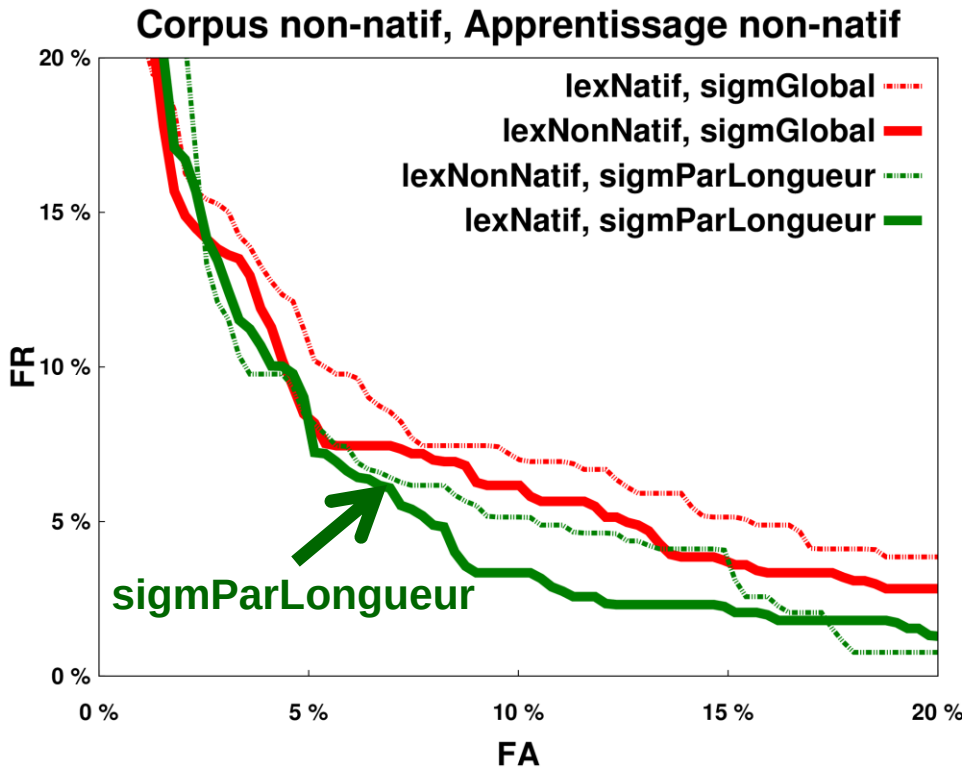
Étude du paramètrage

Loria
Laboratoire lorrain de recherche en informatique et ses applications

- **Étude de l'impact des paramètres de l'approche**
  - fonction de décision globale, ou fonction dépendante de la longueur de l'entrée traitée (courte / moyenne / longue, en fonction du nombre de phonèmes)
  - lexique de prononciations natives ou avec variantes non-natives
  - type des données utilisées pour l'apprentissage des paramètres
  - chaque critère indépendamment ou les trois simultanément

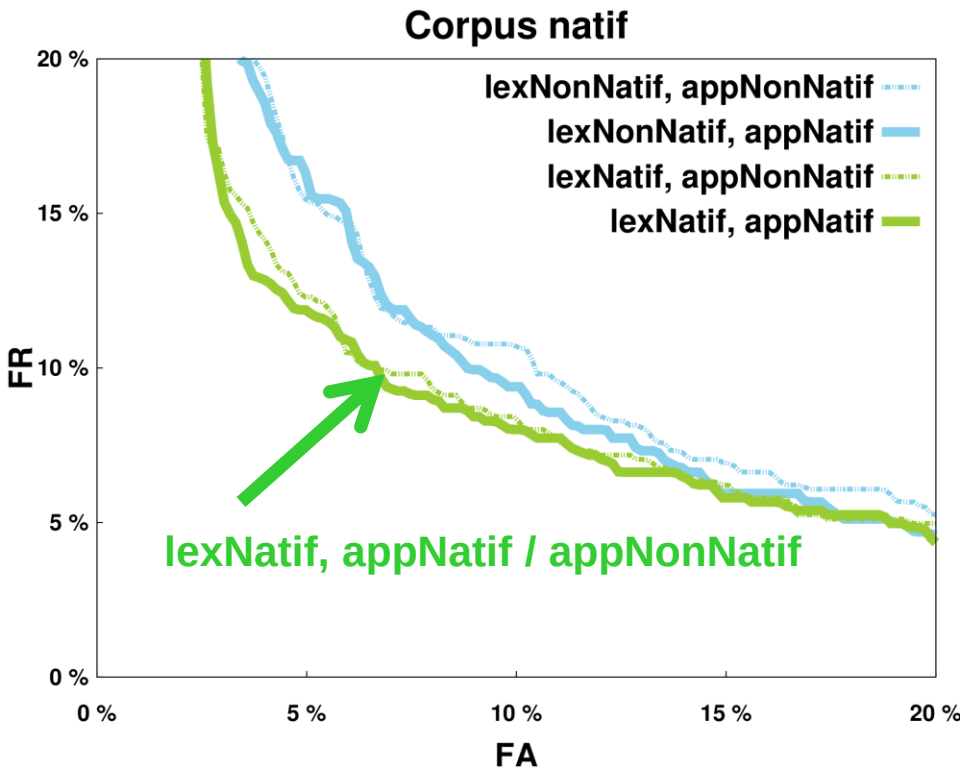
Résultats: fonction de décision

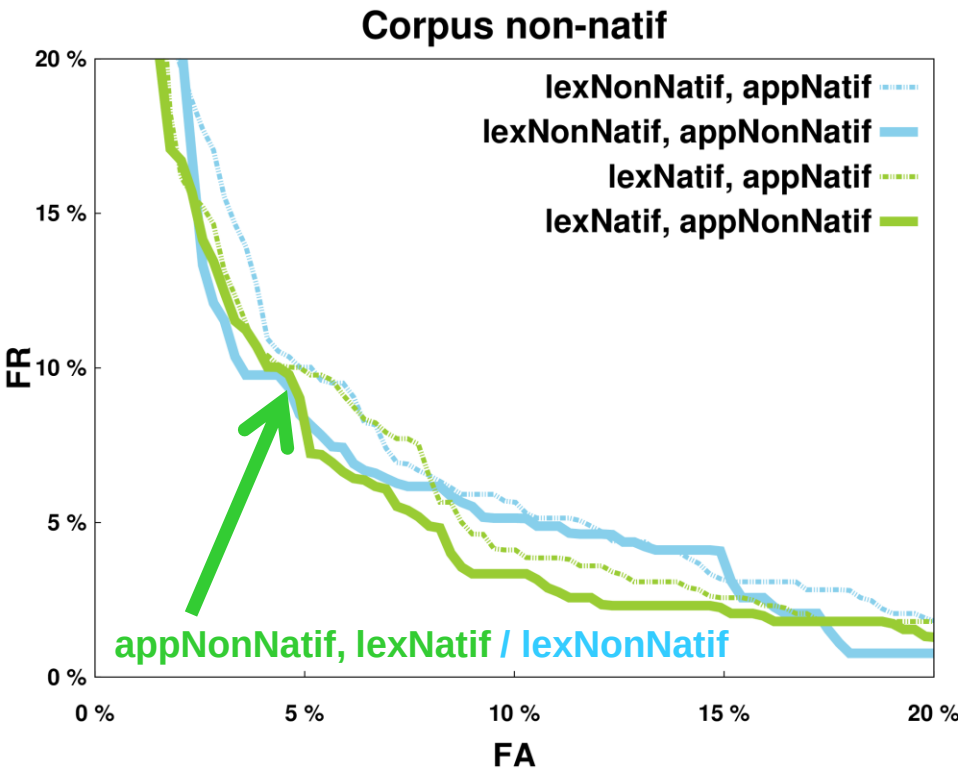




L'utilisation des fonctions dépendantes de la longueur des transcriptions est plus performante

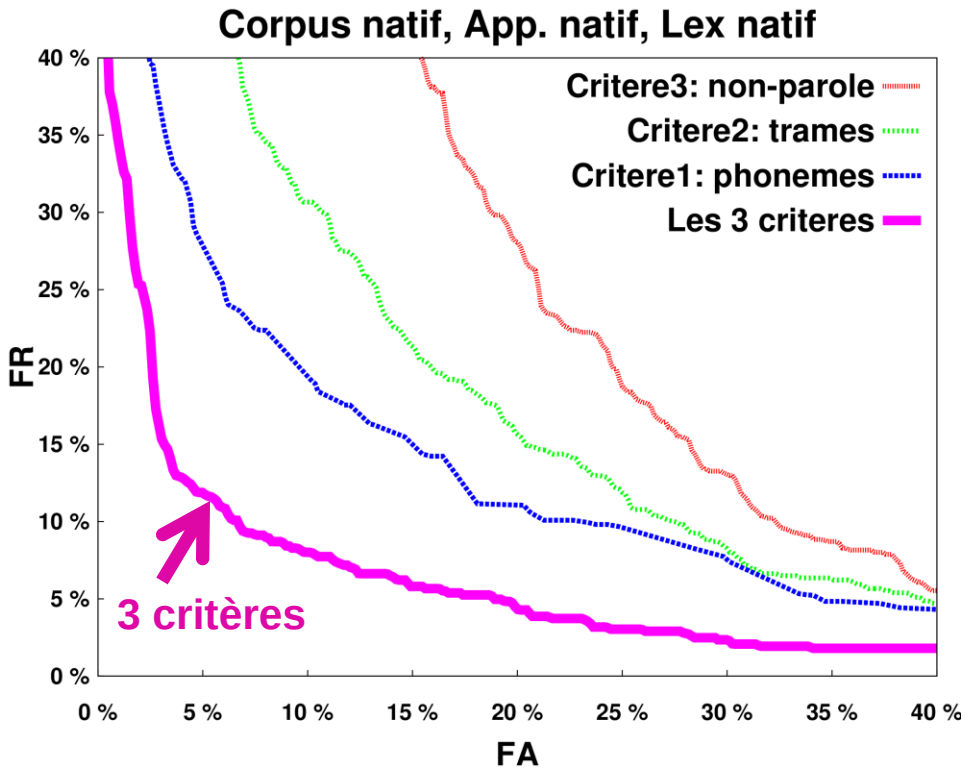
Résultats: lexique et apprentissage

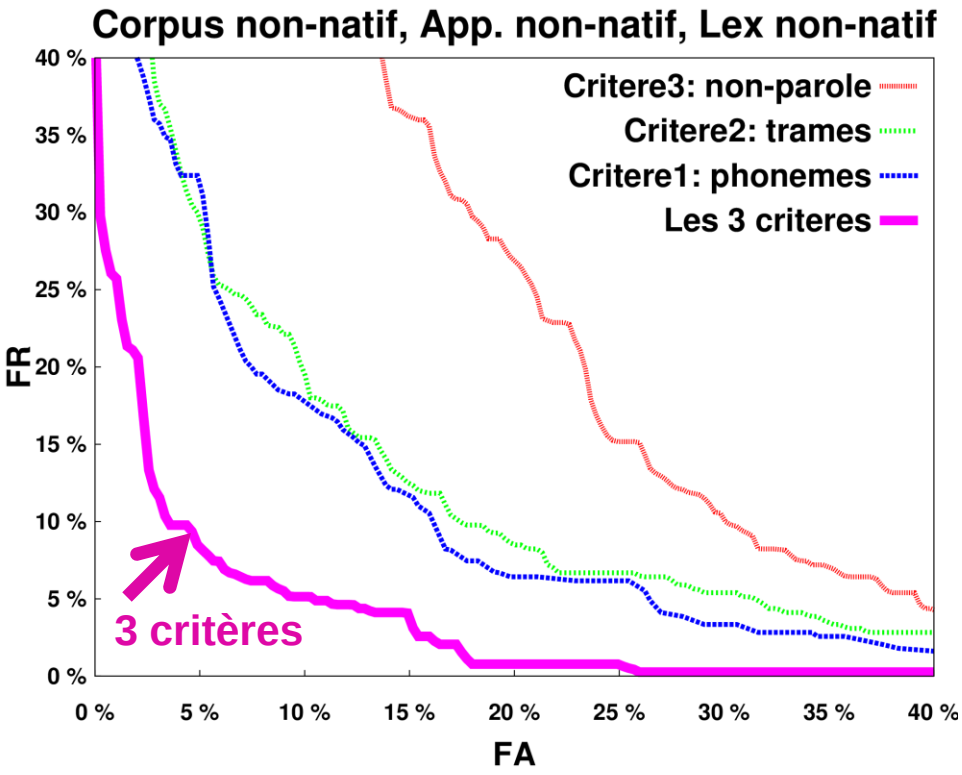




- L'utilisation d'un lexique natif est important pour le corpus natif
  - Pour le corpus non-natif les deux lexiques donnent des résultats similaires
    - Il faut apprendre les fonctions de décision sur le même type des données

Résultats: 1 critère ou 3 critères





Les résultats sont meilleurs si on combine les 3 critères

Résultats

Meilleurs résultats obtenus pour le **corpus non-natif**

Taux de fausse acceptation **4.9%**

Taux de faux rejet **6.7%**

F-mesure **94.2%**

Meilleurs résultats obtenus pour le **corpus natif**

Taux de fausse acceptation **6.4%**

Taux de faux rejet **9.5%**

F-mesure **92.0%**

Plan

Loria
Laboratoire lorrain de recherche en informatique et ses applications

- Contexte et problematique
- Méthodologie
- Expériences et résultats
- **Conclusions**

Conclusions

Afin de rejeter les entrées incorrectes tout en tolérant les défauts de prononciations non-natives il est préférable de :

- Utiliser des fonctions de décision dépendantes de la longueur des transcriptions
- Utiliser des variantes de prononciation natives dans le lexique
- Apprendre les fonctions de décision sur le même type de données
- Utiliser les 3 critères simultanément

Conclusions

Quelques points à aborder:

- Étudier l’impact de l’ajout des prononciations alternatives pour les locuteurs natifs
- Automatiser la génération de variantes de prononciation non-natives

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Merci pour votre attention !

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Questions ?