

# Vers des outils automatiques pour l'évaluation de locuteurs atypiques

*Lionel Fontan<sup>1,2</sup>, Thomas Pellegrini<sup>1</sup>, Jérôme Farinas<sup>1</sup>, Julie Mauclair<sup>1,3</sup>, Vincent Laborde<sup>1</sup>, Halima Sahraoui<sup>4</sup>, Xavier Aumont<sup>2</sup>, Julia Olcoz<sup>5</sup>, Alberto Abad<sup>6,7</sup>*

<sup>1</sup>Université de Toulouse; UPS; IRIT; Toulouse, France

<sup>2</sup>Archean Technologies; Montauban, France

<sup>3</sup>Université Paris Descartes, France

<sup>4</sup>Université de Toulouse; UT2J; OCTOGONE; Toulouse, France

<sup>5</sup>ViVoLAB - Voice Input Voice Output Laboratory; I3A; Universidad de Zaragoza, Zaragoza, Espagne

<sup>6</sup>L2F - Spoken Language Systems Laboratory; INESC-ID; Lisbon, Portugal

<sup>7</sup>IST - Instituto Superior Técnico, Universidade de Lisboa, Portugal

# Plan de la présentation

- ① Introduction : quelles évaluations pour quels objectifs ?
- ② Mesures issues de moteurs de reconnaissance automatique de la parole (RAP)
  - Qu'est-ce que la RAP ?
  - Mesures issues des sorties brutes du moteur de reconnaissance
  - Mesures issues de comparaisons avec des modèles de prononciation
- ③ Exemples d'études
  - Application à la parole perçue par des patients presbyacousiques
  - Application à la parole de locuteurs japonophones en FLE
- ④ Conclusion et perspectives

# Introduction : quelles évaluations pour quels objectifs ?

- Observer qu'un signal de parole soit plus ou moins bien perçu et interprété sert de nombreux objectifs, comme l'évaluation :
  - des capacités d'un locuteur (ex. pathologies de production de la parole, apprentissage d'une L2) ;
  - des capacités d'un auditeur (ex. pathologies d'audition type presbyacousie) ;
  - de canaux de communication (ex. systèmes de sonorisation, salles, radio/télécom)

# Introduction : quelles évaluations pour quels objectifs ?

- Mesures subjectives allant de l'intelligibilité phonétique à la compréhensibilité de messages parlés :
  - % de phonèmes reconnus
  - % de logatomes reconnus (CV, CVC...)
  - % de mots reconnus (monosyllabes, dissyllabes) hors contexte
  - % de mots reconnus dans des phrases / dans un texte
  - % de phrases reconnues
  - % de phrases comprises
  - % de phrases comprises dans un contexte
  - ...

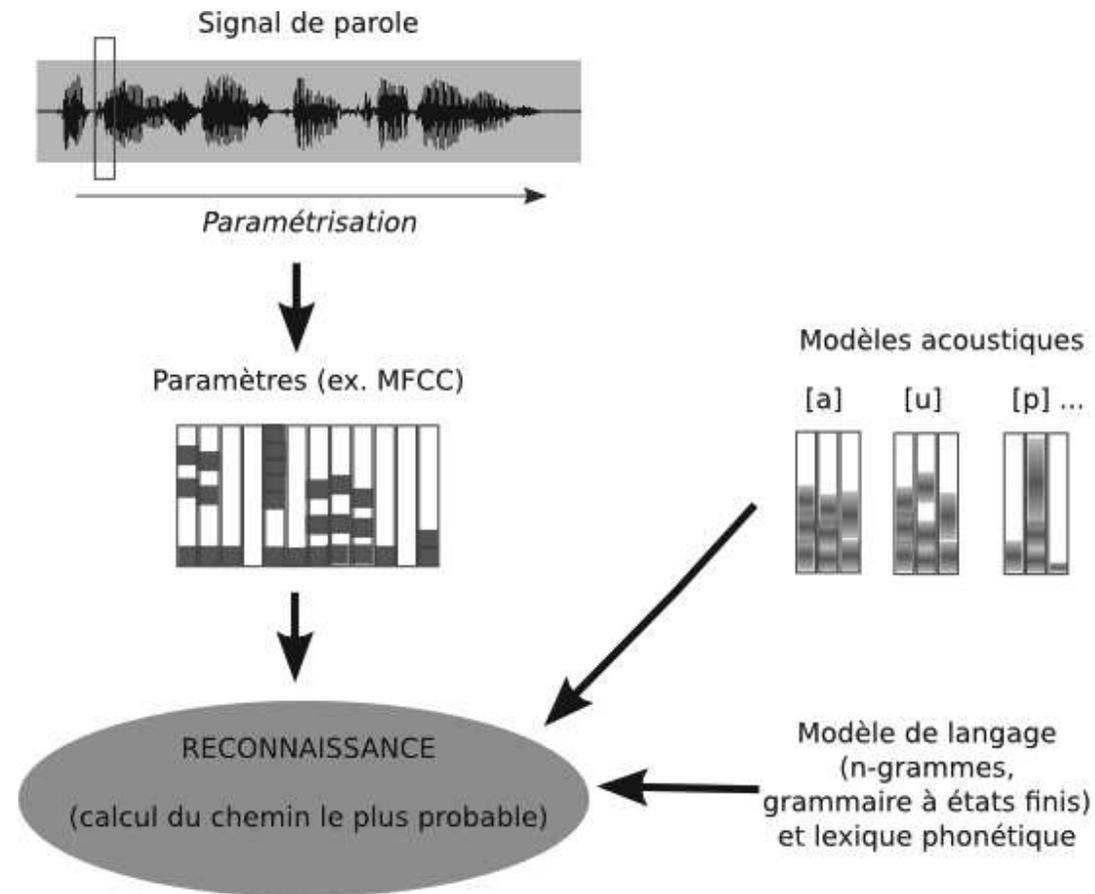
# Introduction : quelles évaluations pour quels objectifs ?

- Choix des mesures subjectives dépendant :
  - de l'aspect qualitatif ou quantitatif de l'évaluation
  - de la priorité accordée à la validité interne (précision, reproductibilité) ou externe (représentativité) des mesures
  - des moyens à disposition (tests de "bas niveau" ou de "haut niveau" souvent difficiles à mettre en œuvre)

→ dans tous les cas, mesures demandant beaucoup de moyens (juges, entraînement des sujets et des juges, statistiques) et subjectives  
→ aujourd'hui matures, les techniques de RAP apparaissent comme des solutions de choix pour l'évaluation de l'intelligibilité et de la compréhension de la parole

# Mesures issues de moteurs de RAP

- Qu'est-ce que la RAP ?



# Mesures issues de moteurs de RAP

- Mesures issues des sorties brutes du moteur de reconnaissance
  - Sortie brute : texte correspondant à la meilleure hypothèse ou bien aux  $n$  meilleures hypothèses
  - Score d'intelligibilité usuel : *word error rate* (WER)

Stimulus	Reconnu	WER
<i>C'est la panne</i> [selapan]	<i>C'est le banni</i> [selbani]	50%

# Mesures issues de moteurs de RAP

- Mesures issues des sorties brutes du moteur de reconnaissance
  - Sortie brute : texte correspondant à la meilleure possibilité ou bien aux  $n$  meilleures possibilités
  - Scores d'intelligibilité plus fins : distances de Levenshtein, éventuellement pondérées (Fontan *et al.*, 2014)

→ Distance de Levenshtein = distance d'édition entre 2 chaînes (nombre d'opérations d'ajout, de substitution, de suppression :

Ex. Distance de 3 entre /selapan/ et /selbani/

→ Distance de Levenshtein pondérée prend en compte le nombre de traits partagés par 2 phonèmes substitués

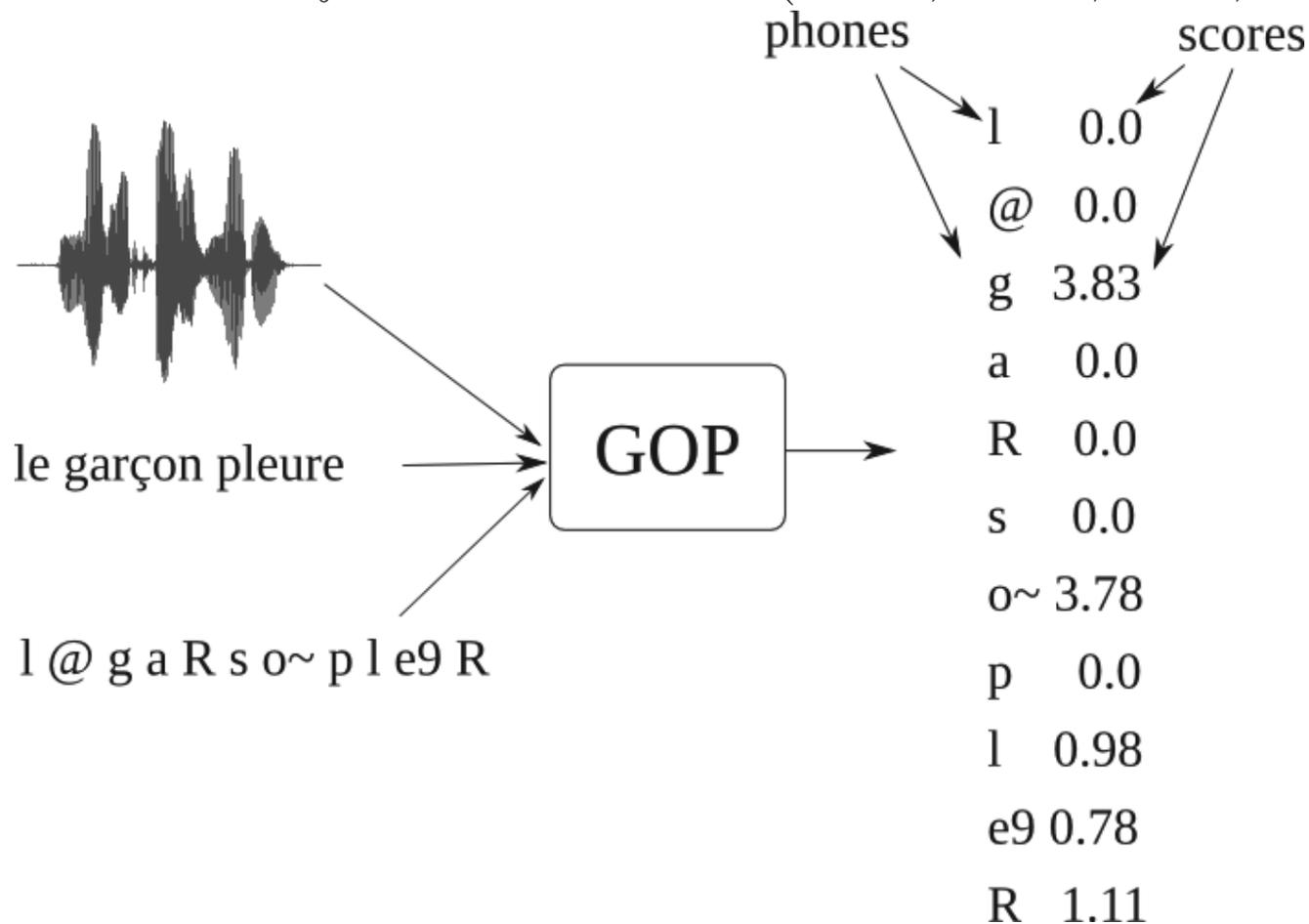
Ex. Distance de 2,125 entre /selapan/ et /selbani/ : 1 (supression de /a/) + 1 (ajout de /i/) + 1/8 (1 trait acoustique sépare /p/ de /b/)

# Mesures issues de moteurs de RAP

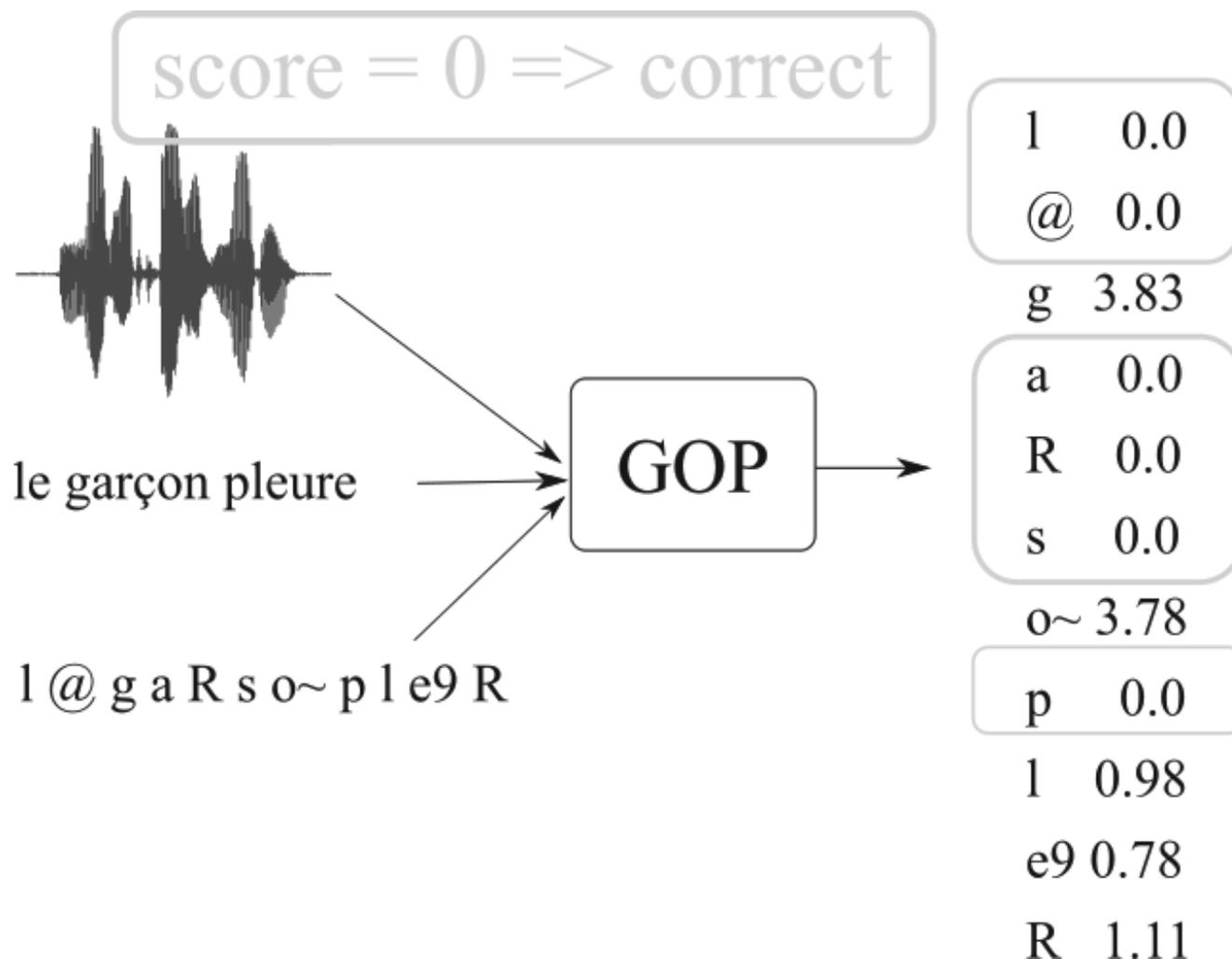
- Mesures issues de comparaisons avec des modèles de prononciation
  - natifs (*Goodness of Pronunciation* – GOP)
  - natifs et non-natifs (*native-likeness*)

# Mesures issues de moteurs de RAP

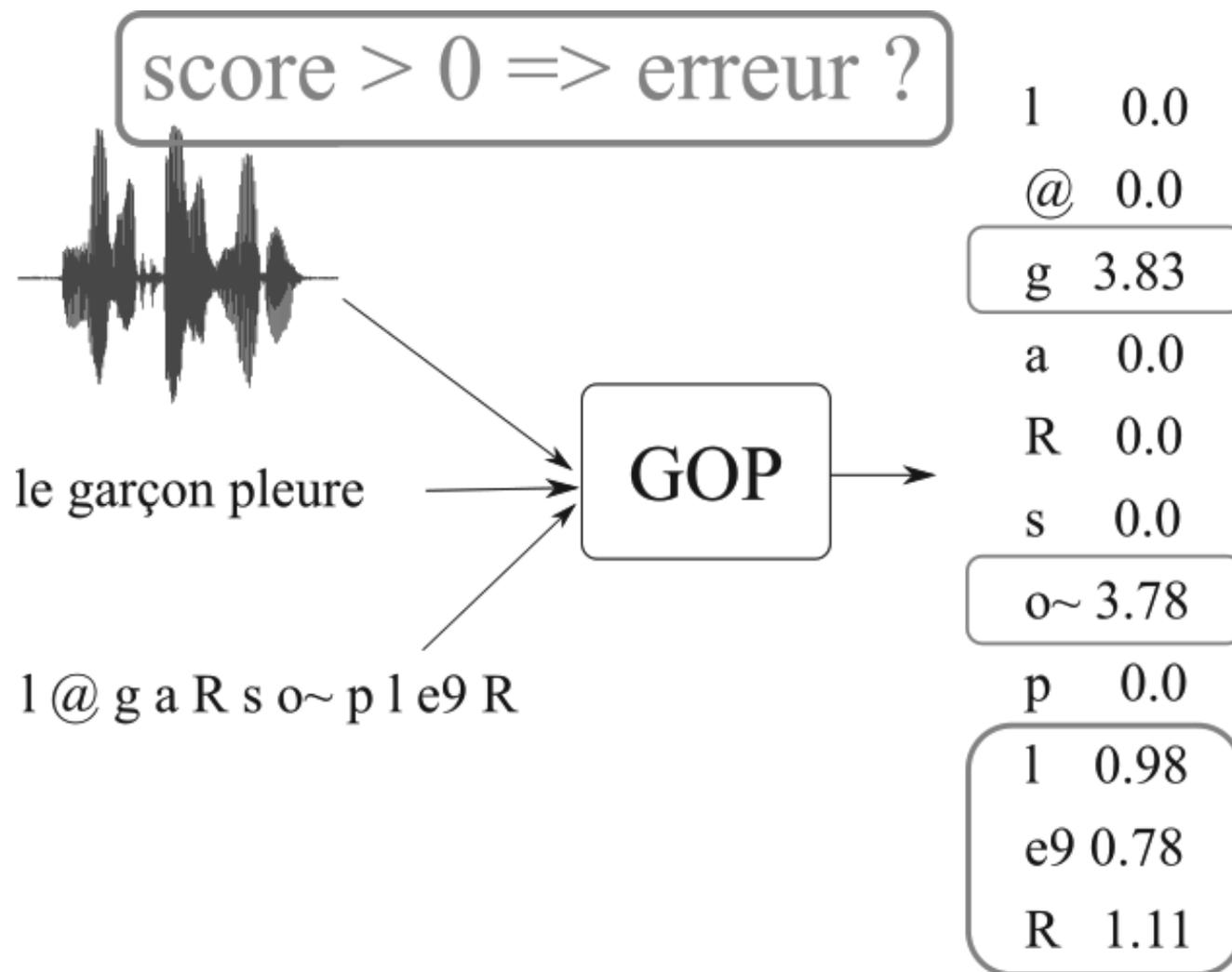
Algorithme *Goodness of Pronunciation* (Witt, 1999; Luo, 2009)



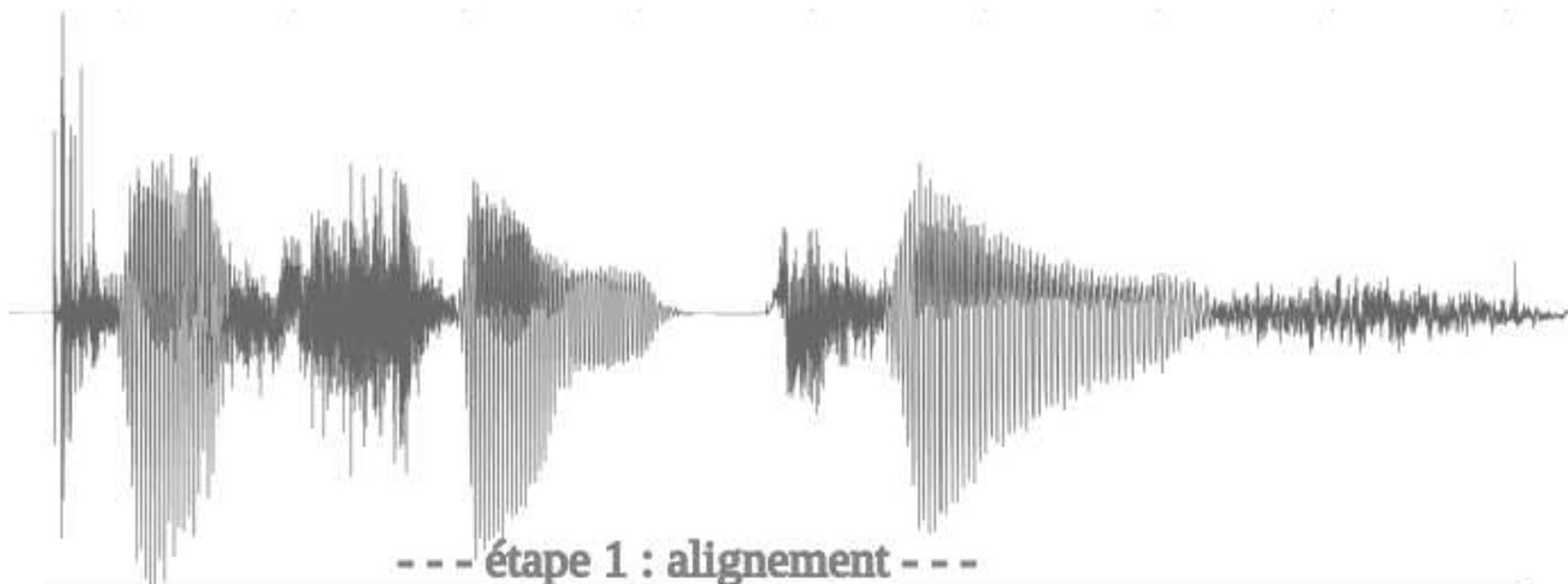
# Mesures issues de moteurs de RAP



# Mesures issues de moteurs de RAP



# Mesures issues de moteurs de RAP



l	g	a	R	s	<b>o~</b>		p	l	e9	R
		...	-702.54	<b>-767.18</b>	...					

--- étape 2 : reconnaissance libre ---

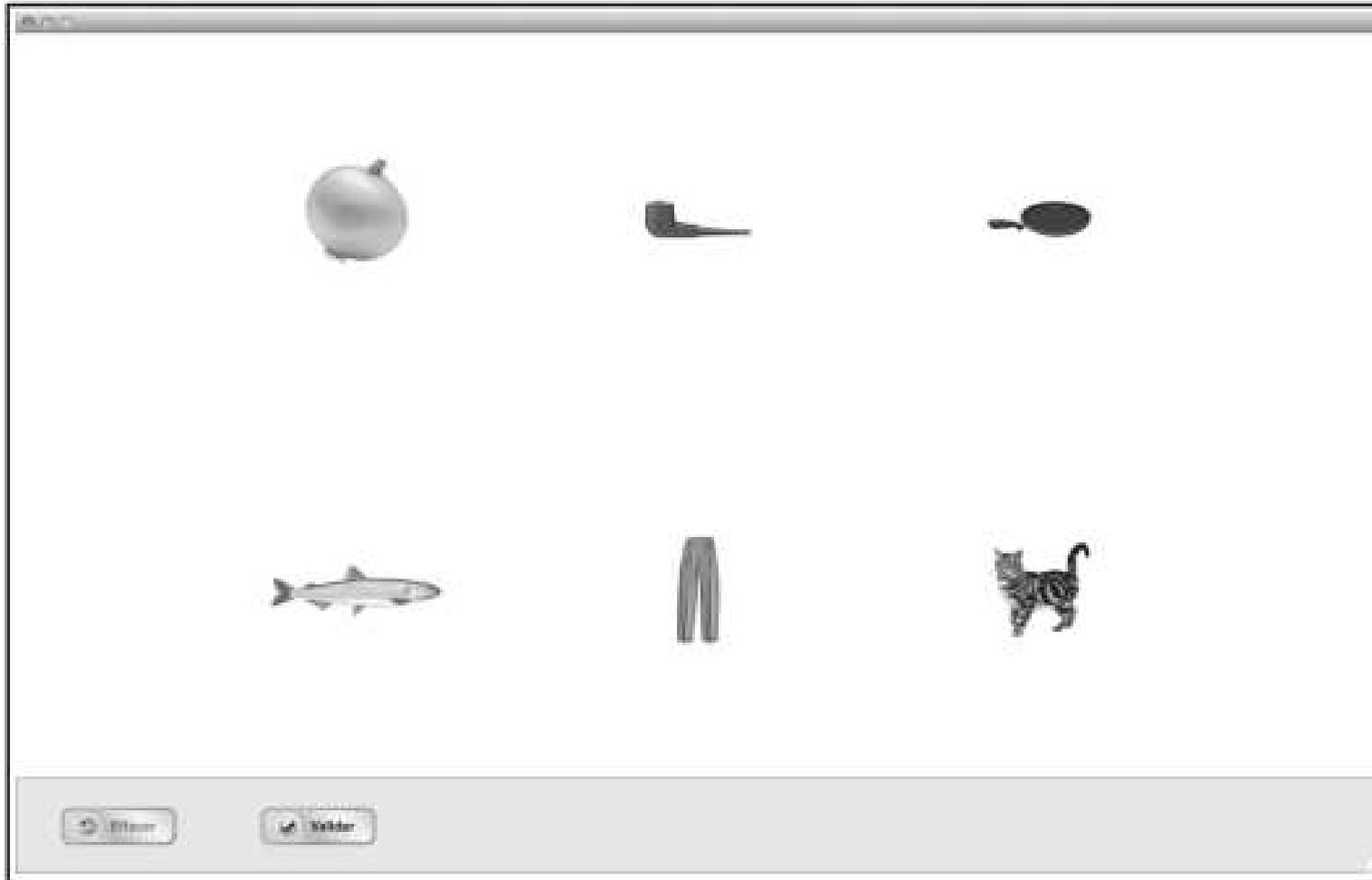
l	k	a	R	s	<b>a~</b>		p	y	e2	U~
		...	-702.54	<b>-725.51</b>	...					

--- étape 3 : calcul des scores GOP ---

$$\text{score}(o\sim) = (-725,51 + 767,18) / 11 = 3,788$$

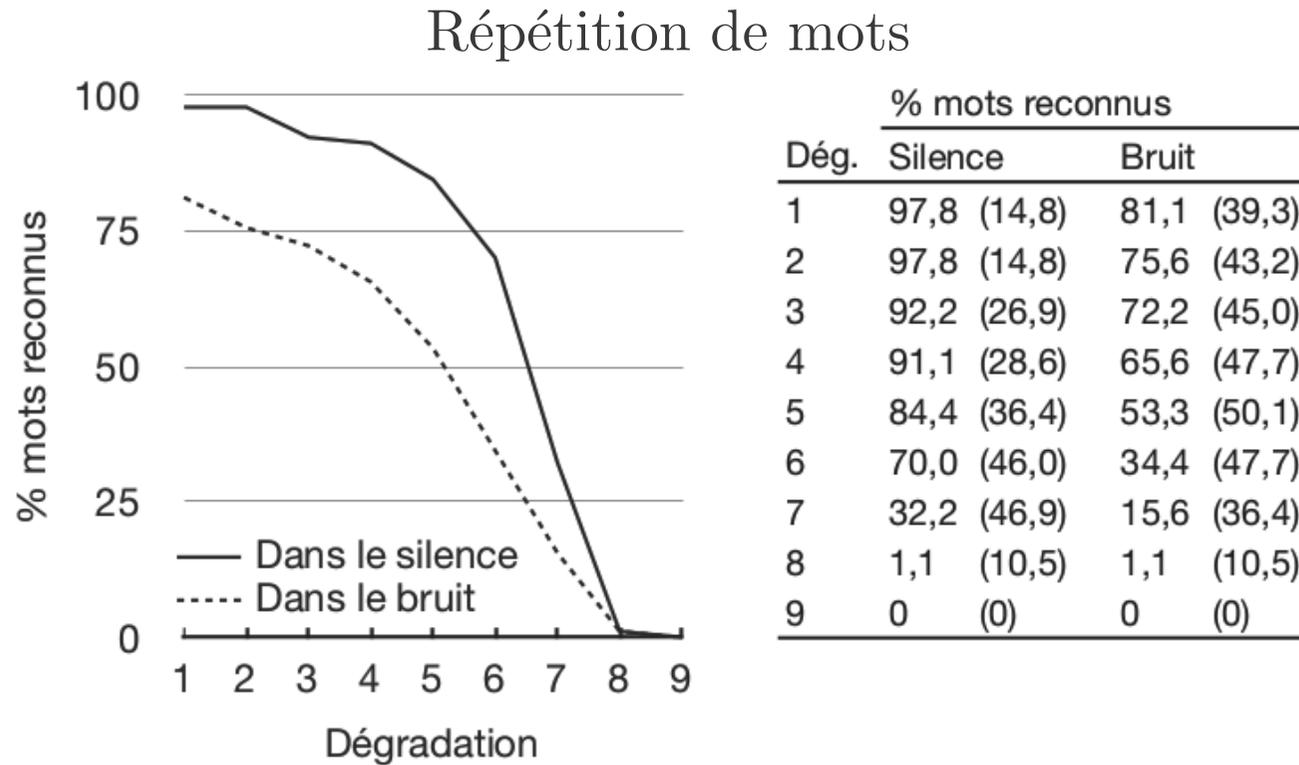
# Exemple d'étude 1 : parole et presbyacousie

- Objectif : prédire des scores allant de l'intelligibilité de mots à la compréhension de phrases, dans le cadre de pertes auditives liées à l'âge
- Méthode :
  - ① Enregistrement d'un corpus de mots et phrases
  - ② Simulation de pertes auditives typiques (60 à 110 ans) par traitement du signal
  - ③ Recueil de scores subjectifs d'intelligibilité (répétition de mots / phrases) et de compréhension de phrases (test EloKanz - Fontan *et al.*, 2013)
  - ④ Tuning d'un moteur de RAP pour coller aux résultats subjectifs



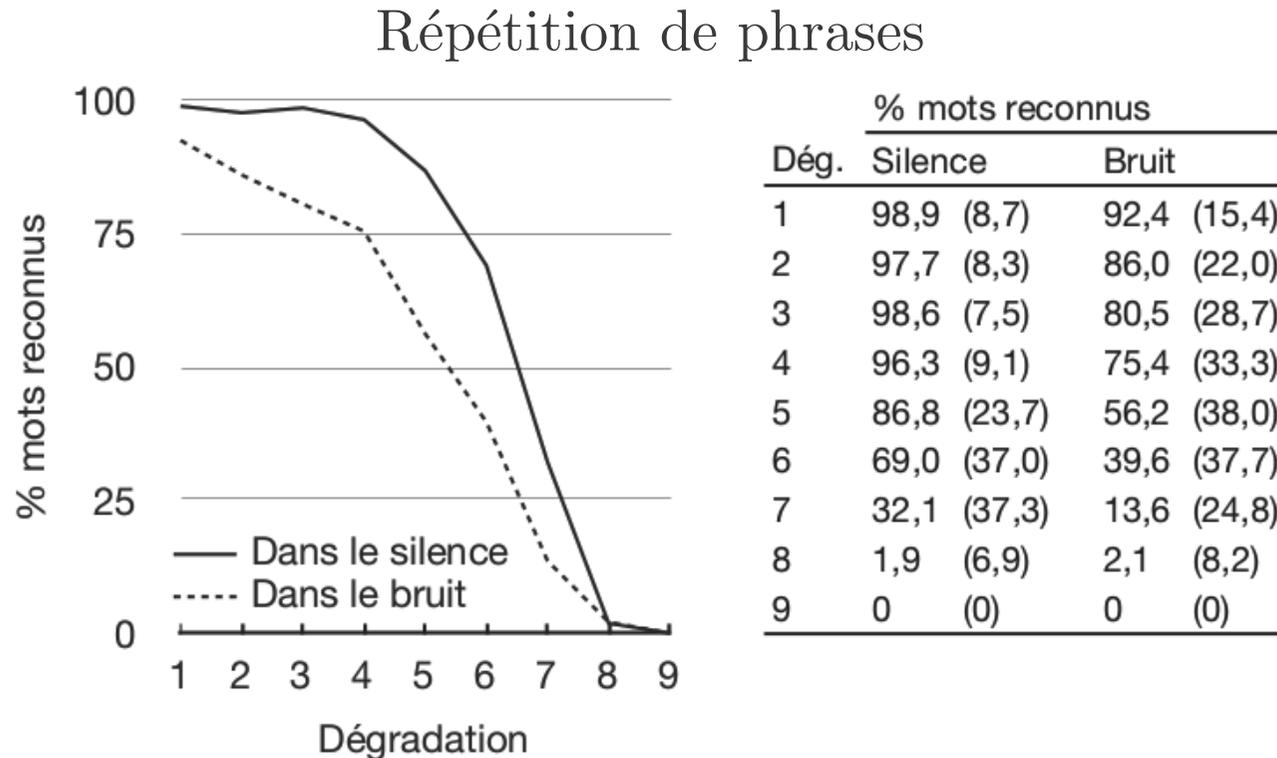
# Exemple d'étude 1 : parole et presbyacousie

- Résultats : scores subjectifs



# Exemple d'étude 1 : parole et presbyacousie

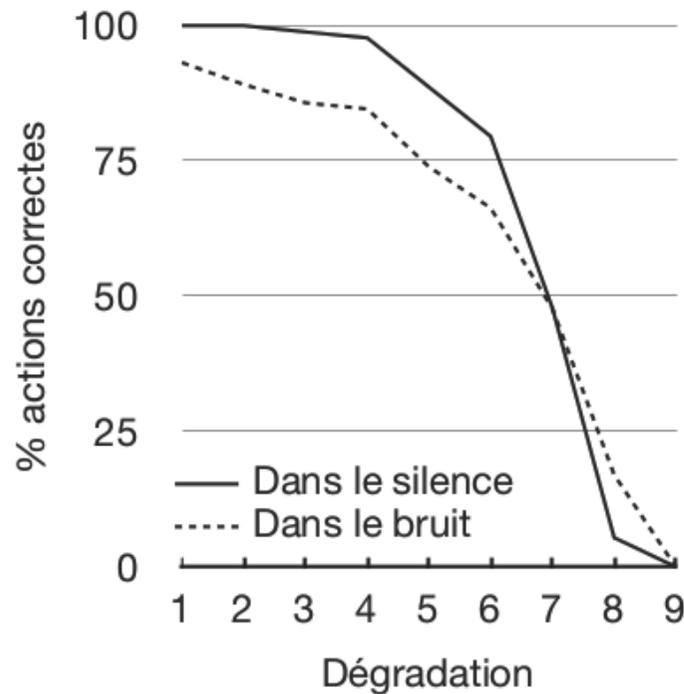
- Résultats : scores subjectifs



# Exemple d'étude 1 : parole et presbyacousie

- Résultats : scores subjectifs

## Compréhension de phrases

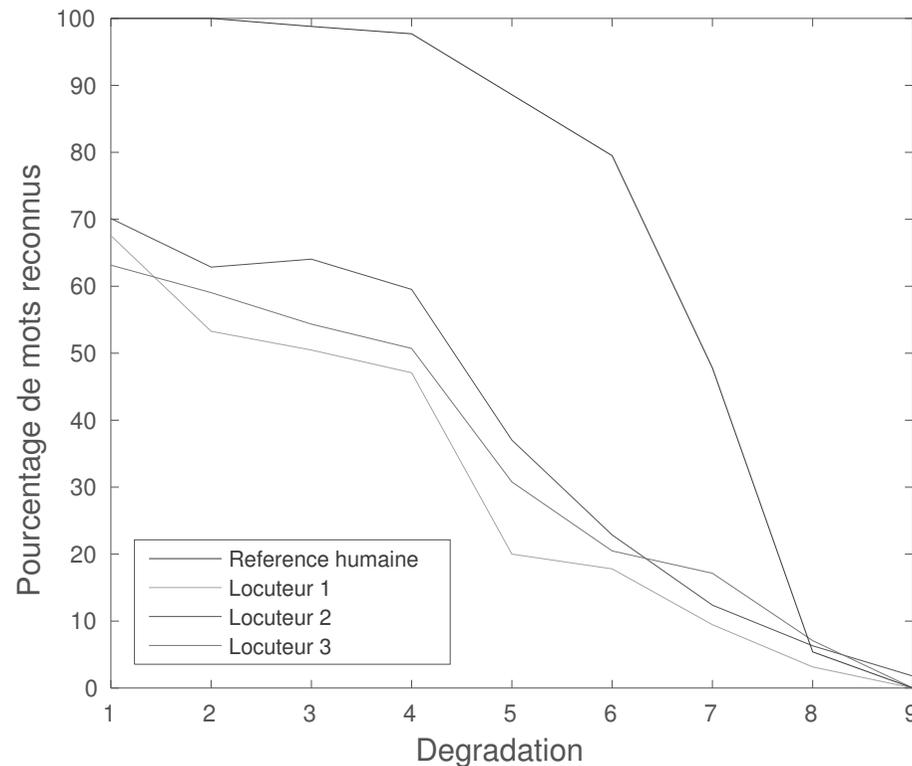


Dég.	% actions correctes	
	Silence	Bruit
1	100 (0)	93,2 (25,4)
2	100 (0)	89,1 (31,3)
3	98,8 (10,7)	85,8 (35,2)
4	97,7 (15)	84,6 (36,3)
5	88,6 (31,9)	73,9 (44,2)
6	79,5 (40,6)	66,3 (47,6)
7	47,8 (50,2)	47,7 (50,2)
8	5,4 (22,8)	17,0 (37,8)
9	0 (0)	0 (0)

# Exemple d'étude 1 : parole et presbyacousie

- Résultats : scores automatiques de compréhension de phrases

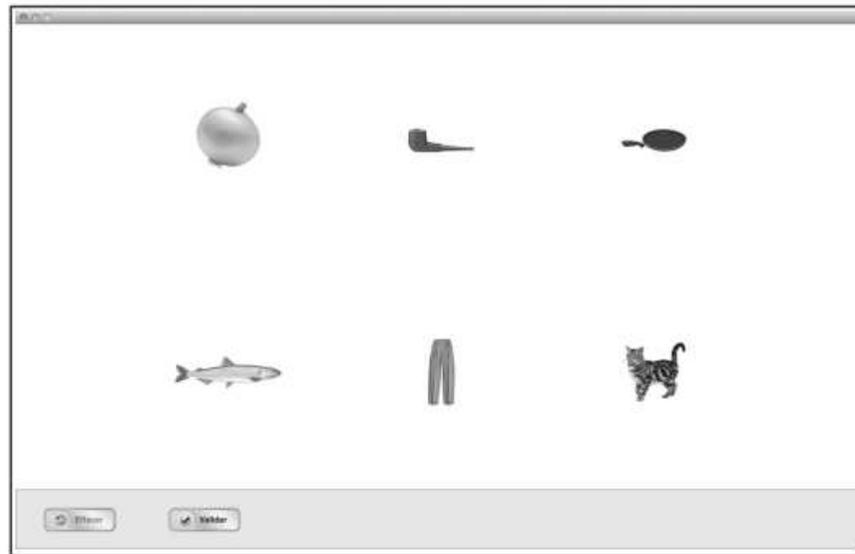
→ avec modèle de langage statistique large vocabulaire : scores plutôt linéaires (corrélation avec scores subjectifs : .87)



# Exemple d'étude 1 : parole et presbyacousie

- Résultats : scores automatiques de compréhension de phrases

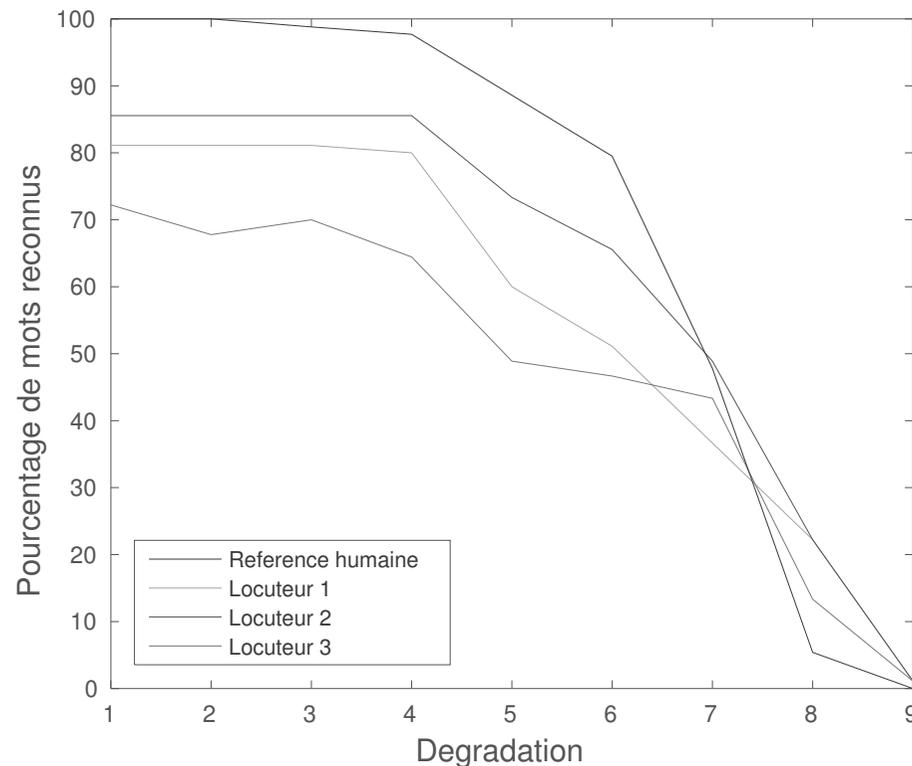
→ idée pour améliorer la corrélation : s'inspirer des effets subjectifs top-down en utilisant une grammaire à états finis (structures syntaxiques et lexique fini)



# Exemple d'étude 1 : parole et presbyacousie

- Résultats : scores automatiques de compréhension de phrases

→ Corrélation obtenue avec scores humains : .99 (locuteur homme)



# Exemple d'étude 2 : locuteurs japonophones

- Objectif : détecter les erreurs de prononciation à l'aide de scores GOP et d'information complémentaire
- Méthode
  - ① Enregistrement d'un corpus de mots et phrases répétés par des apprenants japonophones
  - ② Annotation manuelle des réalisations phonétiques par deux annotateurs (accord : 85%)
  - ③ Détection automatique d'erreurs de prononciation par seuillage des scores GOP (système de base)
  - ④ Prise en compte d'information complémentaire à l'aide d'un classifieur de type régression logistique
  - ⑤ Evaluation et comparaison de performance à l'aide des annotations manuelles

# Exemple d'étude 2 : locuteurs japonophones

- Projet PHON-IM : étudier les changements longitudinaux de la perception et production d'apprenants japonophones en FLE
- Programme d'échange annuel d'étudiants entre l'université Ritsumeikan et l'université Jean Jaurès

## Corpus d'apprentissage : BREF

- 80 locuteurs français natifs
- Environ 66 heures de parole

## Corpus de test : PHON-IM

- 23 locuteurs : 71 mots dissyllabiques, 9 phrases par locuteur
- Environ 1 heure de parole
- Phonèmes d'étude : /R/ et /v/

# Exemple d'étude 2 : locuteurs japonophones

corpus	BREF		PHON-IM	
	correct	incorrect	correct	incorrect
/R/	21K	16K	215	128
/v/	5K	3K	267	50

/R/ prononcé [l]

signal

a l s

alignement

a R s

reconnaissance

a l s

# Exemple d'étude 2 : locuteurs japonophones

- 1 Approche de base : par seuillage
  - seuil(/R/) = 1,13
  - seuil(/v/) = 2,97
- 2 Approche classifieur : score GOP + 6 paramètres

$$p(y = 1|x; \theta) = 1/(1 + \exp(-\theta^T x))$$

## Résultats

Model	baseline	logistic regression	
Features	F-GOP	F-GOP	+1+3+4
Accuracy	63.8%	64.4%	77.1%

# Exemple d'étude 2 : locuteurs japonophones

Meilleure combinaison de paramètres :

- ① Identité du phone reconnu
- ② Nombre de traits distinctifs différents entre le phone attendu et le phone reconnu
- ③ Identité des phones voisins (co-articulation, position)

Détails

- ① /R/ : 55% reconnus [R], substitutions : [f] (13%) et [pause] (9%)
- ② /v/ : 25% reconnus [v], substitutions : [f] (41%), [b] (1%, ?)

# Conclusion et perspectives

- Les techniques de RAP sont aujourd'hui suffisamment matures pour fournir des indices rapides et objectifs de production et de réception de la parole
- Différentes études ont montré une forte corrélation entre scores automatiques et scores subjectifs de perception et de compréhension

→ Nombreuses applications en pratique et en recherche, notamment en acquisition / enseignement des L2

# Conclusion et perspectives

- Néanmoins la prédiction qualitative des (mis)perceptions est encore peu explorée (ex. est-ce qu'un SRAP a aussi tendance à reconnaître un [b] quand un Japonophone réalise un /v/ ?)
- De récents travaux suggèrent que les outils automatiques se rapprochent de la perception humaine du point de vue qualitatif aussi (ex. l'étude de systèmes de réseaux neuronaux profonds de Nagamine *et al.*, 2015), et les recherches se concentrent sur ce sujet (ex. réseau Marie Curie *Inspire*)

→ Pistes futures :

- Inclure des variantes de prononciation non-natives / pathologiques
- Réseaux de neurones profonds (amélioration du GOP)