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

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

[1]Université de Toulouse; UPS; IRIT; Toulouse, France
[2]Archean Technologies; Montauban, France
[3]Université Paris Descartes, France
[4]Université de Toulouse; UT2J; OCTOGONE; Toulouse, France
[5]ViVoLAB - Voice Input Voice Output Laboratory; I3A; Universidad de Zaragoza, Zaragoza, Espagne
[6]L2F - Spoken Language Systems Laboratory; INESC-ID; Lisbon, Portugal
[7]IST - Intituto 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
  - ...

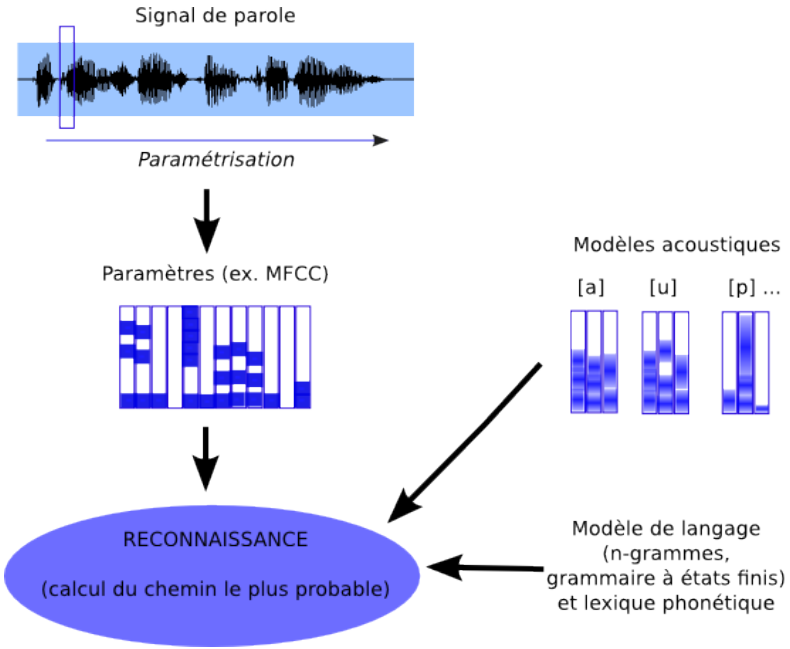
# Introduction : quelles évaluations pour quels objectifs ?

- 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, entrainement 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** |
|---|---|---|
| *C'est la panne* [selapan] | *C'est le banni* [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

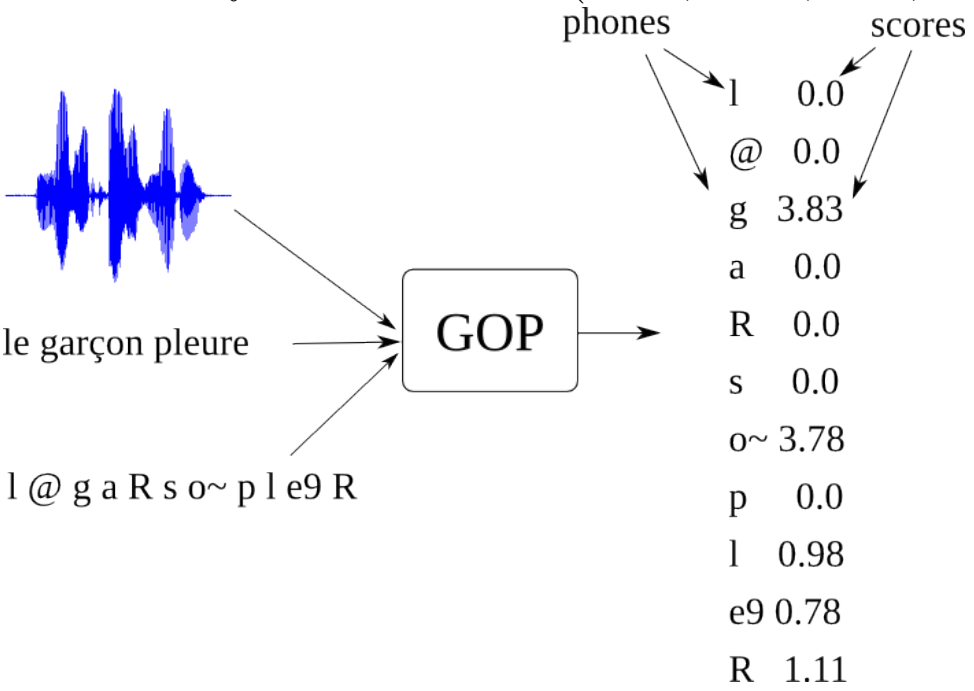
Ex. Distance de 2,125 entre /selapan/ et /selbani/ : 1 (supression de /a/) + 1 (ajout de /i/) + 1/8 (1 trait acoustique sépare /p/ de /b/)

# Mesures issues de moteurs de RAP

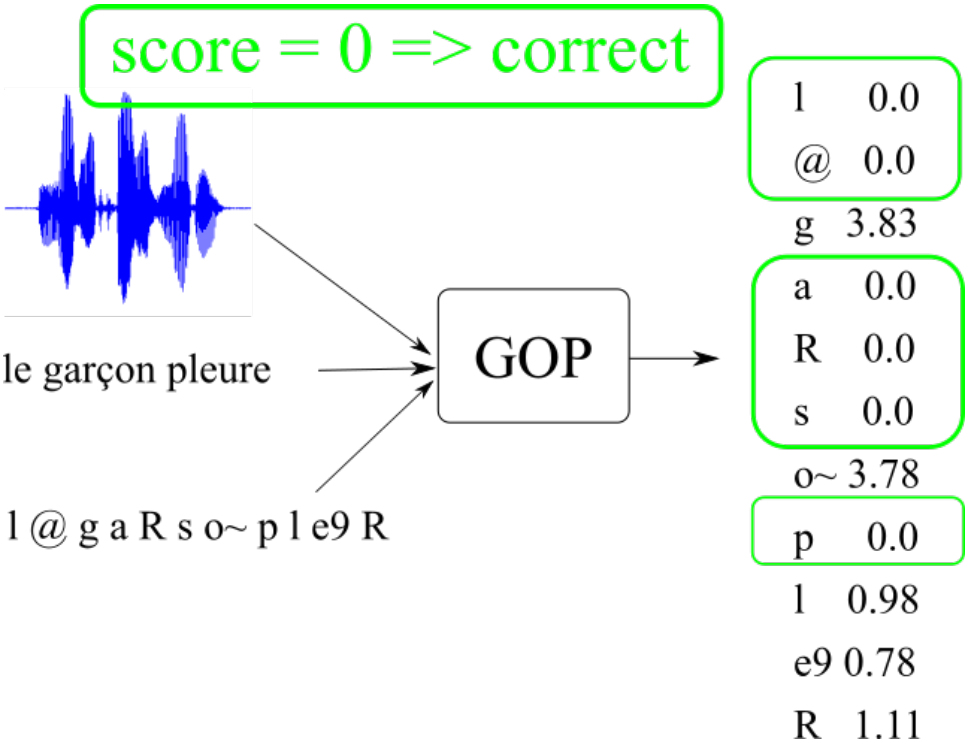
- Mesures issues de comparaisons avec des modèles de prononciation
  - natifs (*Goodness of Pronunciation* – GOP)
  - natifs et non-natifs (*native-likeness*)

# Mesures issues de moteurs de RAP

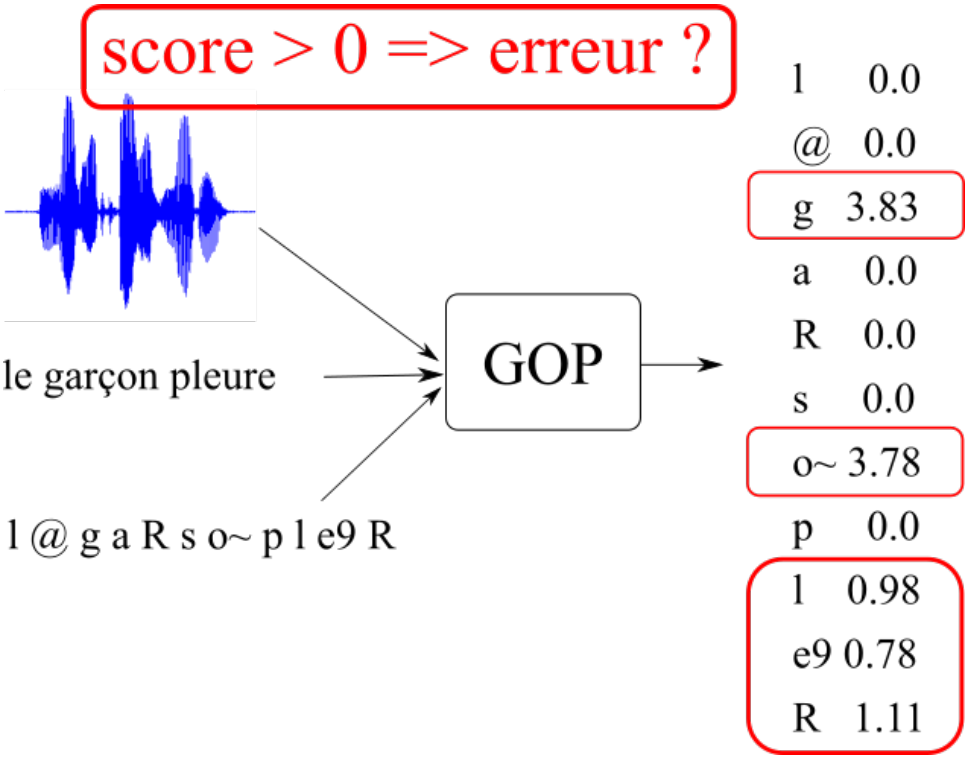
Algorithme *Goodness of Pronunciation* (Witt, 1999; Luo, 2009)



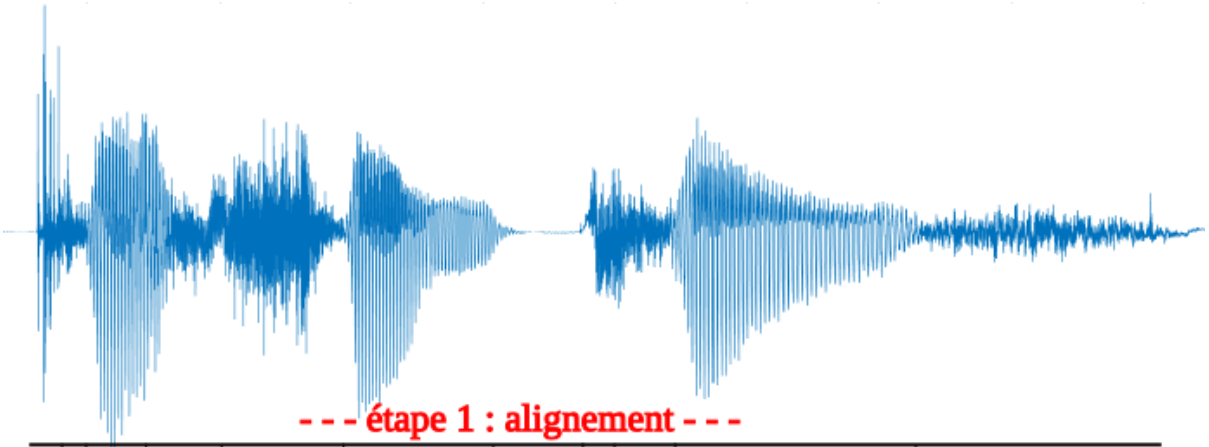
# Mesures issues de moteurs de RAP



# Mesures issues de moteurs de RAP



# Mesures issues de moteurs de RAP



**- - - étape 1 : alignement - - -**

| l | g | a | R | s | o~ | | p | l | e9 | R |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ... | -702.54 | **-767.18** | ... | | | | |

**- - - étape 2 : reconnaissance libre - - -**

| l | k | a | R | s | a~ | | p | y | e2 | U~ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ... | -702.54 | **-725.51** | ... | | | | |

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

**score(o~) = (-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 :

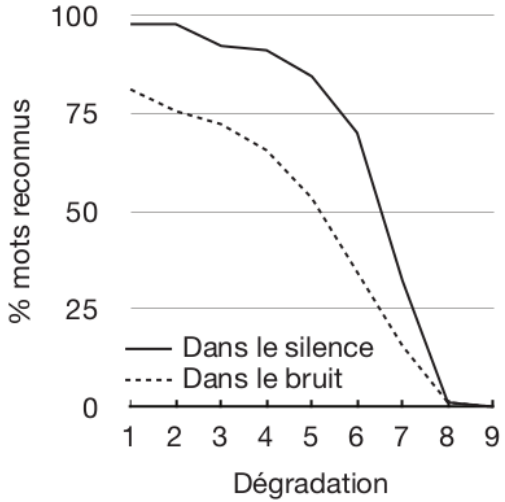
1. Enregistrement d'un corpus de mots et phrases
2. Simulation de pertes auditives typiques (60 à 110 ans) par traitement du signal
3. Recueil de scores subjectifs d'intelligibilité (répétition de mots / phrases) et de compréhension de phrases (test EloKanz - Fontan *et al.*, 2013)
4. Tuning d'un moteur de RAP pour coller aux résultats subjectifs


Effacer
Valider

# Exemple d'étude 1 : parole et presbyacousie

- Résultats : scores subjectifs

Répétition de mots

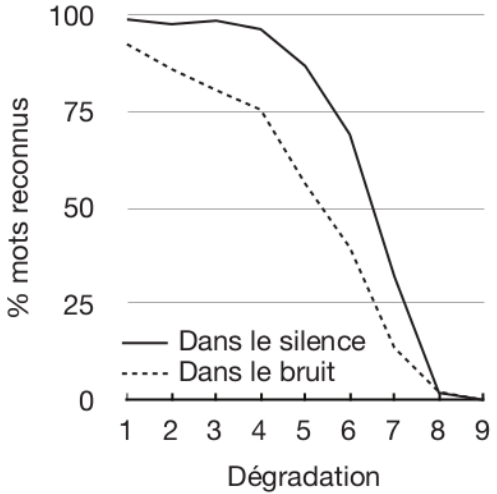



| | % mots reconnus | | | |
|---|---|---|---|---|
| Dég. | Silence | | Bruit | |
| 1 | 97,8 | (14,8) | 81,1 | (39,3) |
| 2 | 97,8 | (14,8) | 75,6 | (43,2) |
| 3 | 92,2 | (26,9) | 72,2 | (45,0) |
| 4 | 91,1 | (28,6) | 65,6 | (47,7) |
| 5 | 84,4 | (36,4) | 53,3 | (50,1) |
| 6 | 70,0 | (46,0) | 34,4 | (47,7) |
| 7 | 32,2 | (46,9) | 15,6 | (36,4) |
| 8 | 1,1 | (10,5) | 1,1 | (10,5) |
| 9 | 0 | (0) | 0 | (0) |

# Exemple d'étude 1 : parole et presbyacousie

- Résultats : scores subjectifs

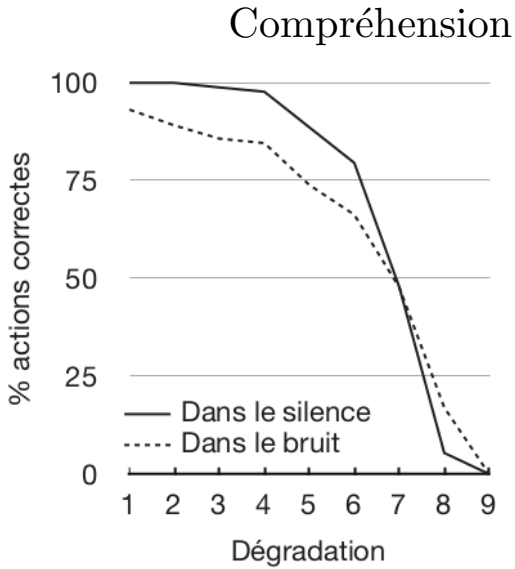
Répétition de phrases




| Dég. | % mots reconnus Silence | | Bruit | |
|---|---|---|---|---|
| 1 | 98,9 | (8,7) | 92,4 | (15,4) |
| 2 | 97,7 | (8,3) | 86,0 | (22,0) |
| 3 | 98,6 | (7,5) | 80,5 | (28,7) |
| 4 | 96,3 | (9,1) | 75,4 | (33,3) |
| 5 | 86,8 | (23,7) | 56,2 | (38,0) |
| 6 | 69,0 | (37,0) | 39,6 | (37,7) |
| 7 | 32,1 | (37,3) | 13,6 | (24,8) |
| 8 | 1,9 | (6,9) | 2,1 | (8,2) |
| 9 | 0 | (0) | 0 | (0) |

# Exemple d'étude 1 : parole et presbyacousie

- Résultats : scores subjectifs

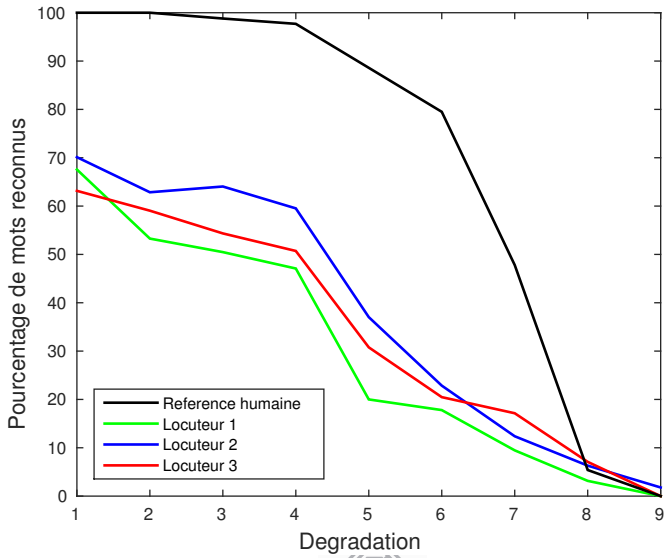



| | % actions correctes | | | |
|---|---|---|---|---|
| Dég. | Silence | | Bruit | |
| 1 | 100 | (0) | 93,2 | (25,4) |
| 2 | 100 | (0) | 89,1 | (31,3) |
| 3 | 98,8 | (10,7) | 85,8 | (35,2) |
| 4 | 97,7 | (15) | 84,6 | (36,3) |
| 5 | 88,6 | (31,9) | 73,9 | (44,2) |
| 6 | 79,5 | (40,6) | 66,3 | (47,6)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

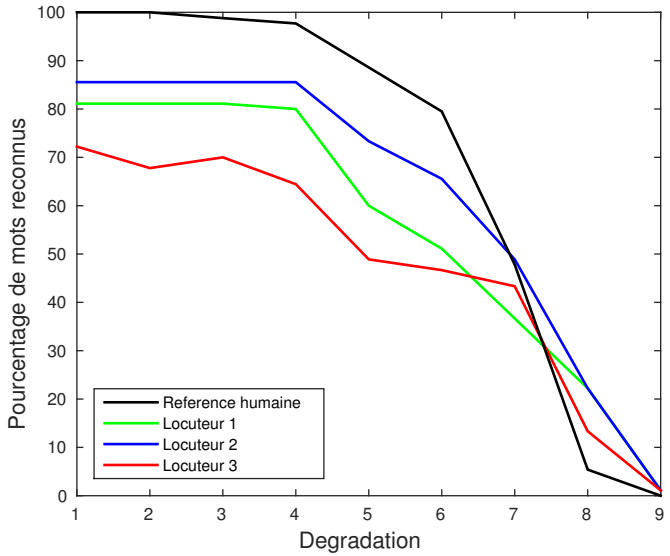
- Résultats : scores automatiques de compréhension de phrases

→ 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

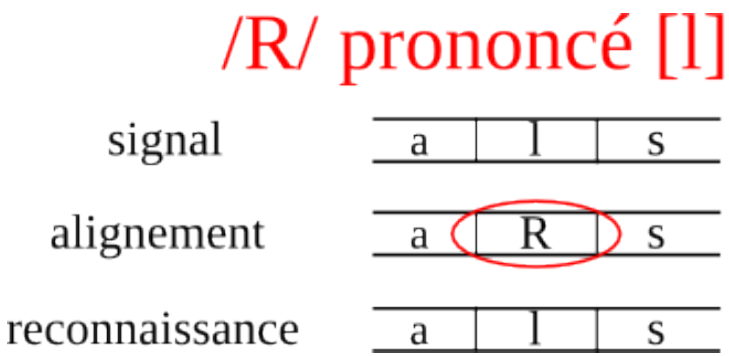
- 80 locuteurs français natifs
- Environ 66 heures de parole

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 |



# 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 ausssi 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)

# Références


- Fontan, L., Farinas, J., Ferrané, I., Pinquier, J., Aumont, X. (2015). Automatic intelligibility measures applied to speech signals simulating age-related hearing loss. In proc. *INTERSPEECH*, Dresden (Germany).
- Fontan L., Gaillard P., Woisard V. (2013). Comprendre et agir : les tests pragmatiques de compréhension de la parole et EloKanz. In R. Sock, B. Vaxelaire, C. Fauth (eds), *La voix et la parole perturbées*, CIPA, Mons, p. 131-144.
- Fontan, L., Magnen, C. Tardieu, J. Ferrané, I. Pinquier, J. Farinas, J. Gaillard, P., Aumont, X. (2014). Comparaison de mesures perceptives et automatiques de l'intelligibilité de la parole : cas de la parole dégradée par une simulation de la presbyacousie. *Traitement Automatique des Langues*, 55(2), 151-174.
- Laborde, V., Pellegrini, T., Fontan, L., Mauclair, J., Sahraoui, H., & Farinas, J. (soumis). Improving the goodness of pronunciation algorithm performance with phonetic features and logistic regression. (*ICASSP 2016*).
- Nagamine, T. , Seltzer, M., (2015). Exploring How Deep Neural Networks Form Phonemic Categories. in Proc. *INTERSPEECH*, Dresden (Germany).