Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information

Vincent Laborde[1], **Thomas Pellegrini**[1], **Lionel Fontan**[2], **Julie Mauclair**[1,3], **Halima Sahraoui**[4], **Jérôme Farinas**[1]

(1) IRIT - Université de Toulouse, 31062, Toulouse, France, (2) Archean Technologies, 1899 av. d'Italie, 82000, Montauban, France

(3) Université Paris Descartes, Paris, France, (4) Octogone-Lordat - Université de Toulouse, 31058, Toulouse, France

thomas.pellegrini@irit.fr, lfontan@archean.fr, sahraoui@univ-tlse2.fr

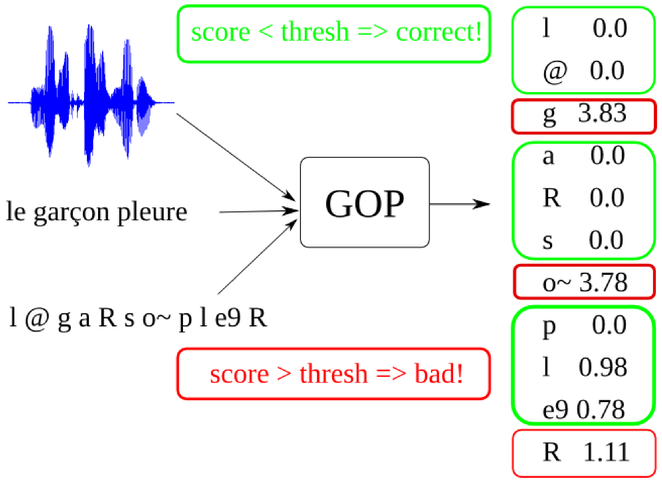
Abstract

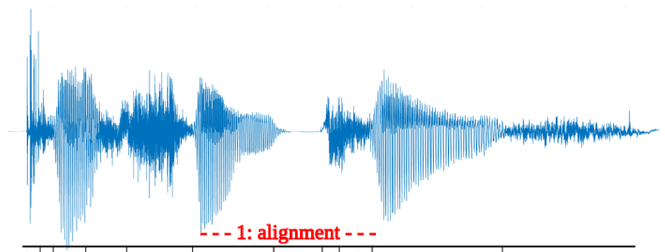
Automatic pronunciation assessment experiments at phone-level
Japanese speakers learning French as a foreign language
Three approaches compared: Goodness Of Pronunciation scores, LDA and logistic regression models
Two typical errors of Japanese speakers were considered: /ʀ/ and /v/
Best method: logistic regression when adding phonetic context information, leading to a 77.1% accuracy

1 Introduction

- Computer-assisted pronunciation training (CAPT) on a second language (L2)
- In this study, segmental level: one score per expected phone
- Two types of approaches:
  - Recognition scores: raw [1], likelihood ratios [3]
  - Classification methods on acoustic features: linear discriminant analysis (LDA) and alike [2]
- High accuracy is key in CAPT
- In this study:
  - Comparison between both methods + use of a classifier
  - Addition of extra phonetic and phonological features

2 Baseline: the forced-GOP algorithm





$$\text{f-GOP}(p) = \left| \log(P(O^p|p) - \max_{j=1..J} \log(P(O^p|p_j)) \right|$$

3 Logistic Regression: f-GOP + LR

- Linear classifier: $\mathrm{P}(y = +1) = 1/(1 + \exp(-\theta x))$
- Input: F-GOP scores and extra features
- Pros: no decision thresholds to set, feature importance given by the weights

4 Speech material

- BREF: read speech corpus recorded from French native speakers, 100 hours, 120 speakers
- Artificial pronunciation errors were introduced in BREF
- PHON-IM: disyllabic words repeated by FSL Japanese native speakers, 1 hour, 23 speakers

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

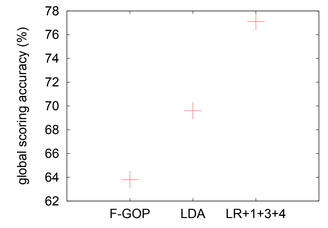
Phoneme	Position		
	initial	*intervocalic*	*final*
/ʀ/	47.3%	50.5%	88.3%
/v/	74.8%	92.6%	88.0%

5 Additional input features

1. Identity of the recognized phone
2. Log-likelihoods of the expected and recognized phones
3. Number of distinctive phonological features that differ between the two phones
4. Identity of the left and right phone neighbors, if any
5. Ratio between the phone duration and the duration of the middle state of the HMM

6 Results

- HTK, context-independent acoustic models — 39 monophones
- Scoring accuracy: $\mathrm{SA} = ((\mathrm{CA} + \mathrm{CR})/(\mathrm{CA} + \mathrm{CR} + \mathrm{FA} + \mathrm{FR})) \times 100$



- Global SA: f-GOP (63.8%) < LDA (69.6%) < best LR (77.1%)
- /R/ SA : LDA (62.4%) < f-GOP (68.5%) < best LR (69.1%)
- /v/ SA: f-GOP (58.7%) < LDA (77.3%) < best LR (85.8%)

- /ʀ/ recognition: 55% as [ʀ], 13% as [f] and 9% as a pause
- /v/ recognition: 25% as [v], 41% as [f], 1% as [b]
- Best extra features: the identity of the recognized phone, the number of distinctive phonological features, the phone neighbor identity

7 Conclusions

- LDA with acoustic features is better than standard f-GOP
- Best performance: f-GOP scores and phonetic / phonological features as input to LR (77% accuracy)
- Future experiments: add acoustic features to LR, deep neural networks

References

[1] B. Sevenster, G. de Krom, and G. Bloothooft. Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs. In *Proc. STiLL*, pages 91–94, Marholmen, 1998.

[2] H. Strik, Khiet P. Truong, F. de Wet, and C. Cucchiarini. Comparing classifiers for pronunciation error detection. In *Proc. INTERSPEECH*, pages 1837–1840, 2007.

[3] S.M. Witt. *Use of Speech Recognition in Computer-Assisted Language Learning.* Phd dissertation, University of Cambridge, Dept. of Engineering, 1999.