INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

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

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

[1]IRIT - Université de Toulouse, Toulouse, France
[2]Université Paris Descartes, Paris, France
[3]Octogone-Lordat - Université de Toulouse, Toulouse, France

laborde.lv@gmail.com, {thomas.pellegrini,lionel.fontan,julie.mauclair}@irit.fr,
sahraoui@univ-tlse2.fr, jerome.farinas@irit.fr

Abstract

In this paper, we report automatic pronunciation assessment experiments at phone-level on a read speech corpus in French, collected from 23 Japanese speakers learning French as a foreign language. We compare the standard approach based on Goodness Of Pronunciation (GOP) scores and phone-specific score thresholds to the use of logistic regressions (LR) models. French native speech corpus, in which artificial pronunciation errors were introduced, was used as training set. Two typical errors of Japanese speakers were considered: /ʁ/ and /v/ often mispronounced as [l] and [b], respectively. The LR classifier achieved a 64.4% accuracy similar to the 63.8% accuracy of the baseline threshold method, when using GOP scores and the expected phone identity as input features only. A significant performance gain of 20.8% relative was obtained by adding phonetic and phonological features as input to the LR model, leading to a 77.1% accuracy. This LR model also outperformed another baseline approach based on linear discriminant models trained on raw f-BANK coefficient features.

Index Terms: Computer-assisted language learning, automatic pronunciation assessment, goodness of pronunciation

1. Introduction

Computer-assisted pronunciation training (CAPT) systems aim at automatically assessing pronunciation to help learners in the acquisition of a second language (L2). For assessment at segmental level, a standard approach consists of assigning a pronunciation score to each expected phone realization [1]. Approaches range from the analysis of raw recognition scores [2], likelihood ratios such as native-likeness and Goodness of Pronunciation (GOP) [3], to the definition of scores derived from classification methods such as linear discriminant analysis (LDA) and alike [4]. In GOP approaches, scores are compared to thresholds to decide whether a realization was close enough to a standard one in order to provide feedback to the user. Recent approaches use deep neural network acoustic models to obtain phone likelihoods [5]. If the algorithm erroneously rejects correct pronunciations too often, users might rapidly give up using the tool [1]. Thus, high accuracy is key in CAPT. In [6], typical error patterns are added as pronunciation variants in the pronunciation lexicon in order to improve the ASR quality for the learners, but no error prediction quantitative evaluation is provided by the authors. Other CAPT systems use low-level acoustic features, such as MFCCs, as input to phone-specific classifiers that take a binary decision about the correctness of a realization. In [7], for example, LDA was shown to slightly outperform the GOP algorithm.

In the current study, we compare the GOP algorithm with LDA and we propose the use of a logistic regression (LR) classifier on top of a GOP algorithm variant, described in Section 2. The evaluation experiments were conducted on a read speech corpus in French, collected from 23 Japanese speakers learning French as a foreign language (FFL). In order to tackle the lack of non-native speech material, we use the same approach as in [7]: a native speech corpus is aligned with a pronunciation lexicon modified by introducing artificial pronunciation errors corresponding to typical errors from the target learners. The alignment system is then forced to align the speech signal with incorrect phone sequences.

Our methodology, covered in Section 3, consisted of comparing the performance of the baseline GOP and LDA approaches with an LR classifier fed with: 1) GOP scores only, 2) GOP scores and additional phonetic and phonological features that give contextual information, such as the identity of the left and right phone neighbors. The use of phonetic context was successfully used in [7] and in pronunciation modeling for disordered speech [8].

2. The GOP and F-GOP algorithms

The baseline GOP algorithm can be decomposed into three steps: 1) forced phone alignment phase, 2) free phone recognition phase and 3) score computation as the difference between log-likelihoods of the two preceding phases for each forced-aligned phone. Scores usually range between 0 and 10, and large scores indicate potential mispronunciations. The forced alignment phase consists of forcing the system to align the speech signal with an expected phone sequence. On the contrary, the free phone recognition phase determines the most likely phone sequence matching the audio input without constraint (free phone loop recognition). The standard approach to decide whether a phone was mispronounced ("reject") or not ("accept"), consists of setting phone-dependent thresholds on a development set.

In this work, we used a variant called forced-aligned GOP (F-GOP). It is exactly the same as the baseline one with the difference that the phone boundaries found during forced alignment constrain the free phone recognition phase. For each aligned phone, a single phone is recognized. In [9], better correlations between GOP and manual scores were found with F-GOP than with baseline GOP in the context of a CALL experiment.

http://dx.doi.org/10.21437/Interspeech.2016-513

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

Table 1: Number of /ʀ/ and /v/ occurrences in BREF and PHON-IM.

3. Methodology

With the GOP algorithms, phone-specific score thresholds need to be set. To do so, one would ideally need a corpus of non-native speech manually annotated at phone-level. As explained in the introduction, the size of such data sets is generally much smaller than the size of a native speech corpus used to train acoustic models for ASR. Thus, common practice consists of introducing artificial pronunciation errors by substituting phone transcriptions in the pronunciation lexicon used during the GOP score computation [10, 7]. We also used this method to benefit from a large French native speech corpus called BREF. Since our target speakers are Japanese native speakers learning French as a foreign Language (FFL), we focused on the two French phonemes /ʀ/ and /v/, which were reported to be very difficult for Japanese speakers [11]. The most frequent confusions occur between /ʀ/ and /l/ [12], and /b/ and /v/ [13, 6]. Thus, every /l/ in the pronunciation lexicon was substituted by /ʀ/ (so the ASR expect a [ʀ] sound and will get an [l] in the audio), and similarly every /b/ was changed as a /v/. For each target phone, a threshold was calculated by stacking all its F-GOP scores in a single vector, ordered by increasing score, and by searching the threshold that minimized the number of errors equaled to the sum of false accepts and false rejects. In this experience, the thresholds were 1.13 and 2.97 for /ʀ/ and /v/, respectively.

The objective of this work was to improve the baseline GOP and LDA approaches. To do so, we added information to single F-GOP scores in the form of additional features given as entry to a probabilistic model, a logistic regression model (LR). Very popular in particular in natural language processing, this technique is known to obtain performances comparable to support vector machines [14]. Compared to LDA, LR also has the advantage that a single model can be used to evaluate several target phones. We trained LR classifiers on the same corpus on which the thresholds were set for the baseline method (BREF with artificial errors), which is also the case of the two LDA models needed for the two target phones. The LR model weights provide information about the relative importance of the input features. The estimated weight of the GOP score feature was -0.633, a negative value that corresponds to the fact that the larger the GOP score, the more likely a pronunciation error. Weights for the categorical phone identity were 0.627 and 0.445 for /v/ and /ʀ/, respectively. The /v/ weight is slightly larger than the /ʀ/ one, which is also consistent with the fact that the corresponding GOP threshold is higher for that phone.

Results were then compared on a test corpus comprised of read speech collected from an homogeneous group of FFL Japanese students. The /ʀ/ and /v/ realizations were manually labeled as correctly or incorrectly pronounced by two annotators with a solid background in phonetics and experience in transcribing speech in the context of FFL teaching. A high inter-annotator rate of 84.4% showed large consensus in their annotation, with a larger agreement on the /v/ than on the /ʀ/ realizations: 86.1% and 82.9%, respectively. Only the phones for which the annotators agreed on were used for test. Performance is assessed through precision, recall and F-measure of correctly accepted (CA) and correctly rejected (CR) realizations [7]. A scoring accuracy computed as $\mathrm{SA} = ((\mathrm{CA}+\mathrm{CR})/(\mathrm{CA}+\mathrm{CR}+\mathrm{FA}+\mathrm{FR})) \times 100$ was used as a global performance measure, with FA and FR being false accepts and false rejects, respectively.

3.1. Speech material

3.1.1. BREF

The BREF corpus is a read speech corpus recorded from French native speakers. It was designed to provide enough read speech data for the development and evaluation of continuous speech recognition systems in French [15]. It contains over 100 hours of speech material from 120 speakers. All the recorded texts come from the French newspaper *Le Monde*, which correspond to over 20K words and a wide range of phonetic environments (over 300K phones). In this study, a subset comprised of speech from 80 speakers was used. Table 1 shows the number of /ʀ/ and /v/ realizations in the subset: 21K and 5K, respectively. These correspond to true realizations of these two phonemes, thus considered as "correct" pronunciations. Furthermore, 16K of /l/ and and 5K of /b/ realizations were artificially substituted by /ʀ/ and /v/, respectively, corresponding to incorrect realizations of these two last phonemes.

3.1.2. PHON-IM

The PHON-IM project aims at studying the longitudinal changes within the perception and production skills of FSL Japanese native speakers. PHON-IM takes place within the framework of a yearly student exchange program between the Ritsumeikan University (Kyoto, Japan) and Jean Jaurès University (Toulouse, France) [16]. The PHON-IM Japanese learners constitute a rather homogeneous group with a generally low proficiency level in French. Once a year, they come to Toulouse, to learn French in a one-month intensive course, consisting in both general classes and phonetic training classes (perception and pronunciation exercises). To create the corpus used in the current study, 23 speakers were recorded at the beginning and at the end of their stay. They had to listen and repeat 71 disyllabic words or pseudo-words during two sessions, resulting in 58 minutes of recording. Those words and sentences contained the two target phonemes of interest /ʀ/ and /v/. The phone realizations were manually annotated following the procedure we described above. A total of 414 /ʀ/ and 368 /v/ realizations were labeled. On the right-hand side of Table 1 (PHON-IM), the numbers of correct and incorrect labeled instances are given, after selecting the ones which were given the same label by both annotators that totals 82.9% and 86.1% of the occurrences of /ʀ/ and /v/, respectively.

Model	baseline	LDA	logistic regression						
Features	F-GOP	f-BANK	F-GOP	+1	+2	+3	+4	+5	+1+3+4
SA	68.5/58.7	62.4/77.3	71.1/57.1	68.5/81.4	69.1/54.9	69.7/63.7	73.2/57.1	70.8/57.4	69.1/85.8
precisionCA	73.2/91.5	66.0/86.0	71.6/92.3	71.3/91.6	70.9/92.5	72.7/92.7	72.7/92.3	71.4/91.8	69.8/91.7
recallCA	78.6/56.2	82.3/87.3	89.3/53.6	83.3/85.8	86.0/50.6	82.8/61.8	91.6/53.6	89.3/54.3	89.3/91.4
FmeasureCA	75.8/69.6	73.3/86.6	79.5/67.8	76.8/88.6	77.7/65.4	77.4/74.2	81.1/67.8	79.4/68.2	78.4/91.6
precisionCR	58.9/23.5	49.3/26.1	69.3/23.5	60.9/43.3	63.4/22.8	62.2/26.6	75.0/23.5	68.9/23.3	66.2/54.9
recallCR	51.6/72.0	28.9/24.0	40.6/76.0	43.8/58.0	40.6/78.0	47.7/74.0	42.2/76.0	39.8/74.0	35.2/56.0
FmeasureCR	55.0/35.4	36.4/25.0	51.2/35.9	51.0/49.6	49.5/35.3	54.0/39.1	54.0/35.9	50.5/35.4	45.9/55.4

Table 2: Results on the PHON-IM test corpus. In each cell, percentages for /ʁ/ and /v/ are given.

3.2. ASR system setup

As they have been found to be more suitable for CALL applications [17], context-independent acoustic models (39 monophones) were used. This work was carried out with HTK [18]. The acoustic models are three-state left-to-right HMMs with 32 Gaussian mixture components trained on the ESTER corpus [19]. The training corpus is composed of 31 hours of broadcast news clean speech from several French national radio programs. Initialization of models was done with automatic alignments of the Phase I training corpus [20] using Baum-Welch re-estimation. Twelve MFCCs, normalized energy, delta, and delta delta were used as features extracted on 16ms windows with half overlap. These acoustic models are available online [21].

3.3. Additional input features

The F-GOP score and the identity of the expected phone were the baseline features fed to a baseline LR classifier. This configuration is comparable to the one of the threshold-based baseline F-GOP approach, and it allows to observe the impact of using the logistic function instead of using raw thresholds.

For each phone realization, in addition to these two baseline features, five features were computed in order to improve the detection of mispronunciations. All the combinations of the two baseline features and the five extra ones were tested:

1. the identity of the recognized phone, which was expected to be informative since the decoder likelihood ranges depend on the phone identities,
2. the log-likelihoods of the expected and recognized phones, for the same reason as above,
3. the number of distinctive phonological features that differ between the two phones, with the idea that the further the recognized and aligned phones in terms of phonetic properties are, the more probable the mispronunciation is,
4. the identity of the left and right phone neighbors, if any, with the rationale that context matters in pronunciation realization (co-articulation effects),
5. the ratio between the phone duration and the duration of the middle state of the HMM, which is supposed to be the stable and longest state.

4. Results

4.1. Observed articulatory deviances

Table 3 shows the proportion of phones that were labeled as correct realizations of target phonemes by both annotators. As can be seen, the three positions initial, intervocalic and final do not imply the same pattern of performances for the two French phoneme realizations. For example Japanese learners seem to have less difficulty in producing [v] in the intervocalic context, whereas the production of [ʁ] appears to be less problematic in the final position.

This effect is statistically significant: a linear mixed model analysis showed that both the target phoneme ($F(648;1) = 52.3$), position ($F(648;2) = 26.4$) and the interaction target phone * position ($F(648;2) = 15.0$) were highly significant ($P < .001$).

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

Table 3: Phoneme realizations labeled as acceptable by both annotators, as a function of intraword phone position.

The fact that phone position in words may be more or less facilitating for the production of [ʁ] and [v] by Japanese learners of French is well known [22, 23]. For example in the Japanese phonological system the fricative bilabial [β], which is close to [v], is an intervocalic allophone of /b/ in Japanese, which may explain why Japanese learners have less difficulties for producing [v] in this position [24].

4.2. Performance analysis

Table 2 shows the performance results obtained with the baseline F-GOP and LDA approaches, and with the different LR models, when using the F-GOP scores and the identity of the expected phone only (F-GOP column), and when adding each of the five extra features one at a time. The last column gives the results of the best feature combination. In each cell of the table, two numbers are given for /ʁ/ and /v/, respectively.

Figure 1 shows the global scoring accuracy (gSA) obtained with F-GOP, LDA, and the best LR model. The F-GOP approach gave a 63.8% accuracy. The corresponding LR model (second F-GOP column) gave a similar performance of 64.4%.

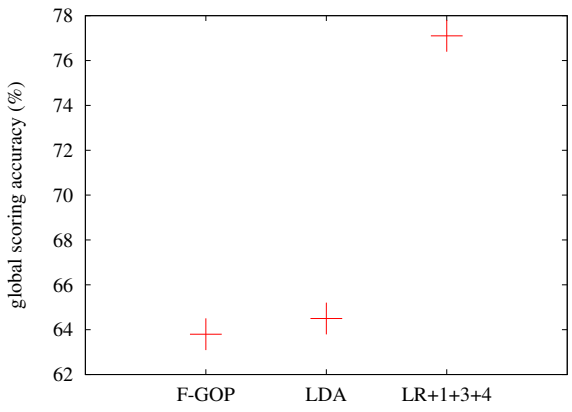

Figure 1: Global scoring accuracy for three systems: baseline F-GOP, LDA, and the best LR system.

By analyzing the results for /ʀ/ and /v/ separately, it appeared that when the recognized phone matches the expected one, then both systems always predict as correct the pronunciations. Fifty-five percent of the 343 expected realizations of /ʀ/ were recognized as [ʀ], and the most frequent substitutions involved [f] (13%) and the model for pauses (9%). This was consistent with the manual annotations, which showed that /ʀ/ realizations were most often transcribed using the Japanese phone [h] – an unvoiced, grave and fricative consonant rather close to [f] or to a breathing pause. For /v/, 25% and 41% of the occurrences were recognized as [v] and [f], respectively. Only 1% of the occurrences were recognized as [b], which is in contradiction with the manual data: [b] was the most frequent alternative phone that the annotators used to transcribe Japanese speakers' productions.

The LDA models outperformed F-GOP and the F-GOP-based LR model for [v] with a 77.3% SA value. It suggests that pertinent information is contained in the raw signal that is well captured by LDA and that is not reflected in ASR likelihoods used to derive the GOP scores. On the contrary, LDA performed slightly worse for [ʀ], with a 62.4% SA value.

Regarding the LR models with extended input features, significant performance improvements were obtained. Adding the identity of the recognized phone (*+1*), yielded to large gains in F-measure: 9.0% and 7.0% absolute for CA and CR, respectively. These improvements impacted the [v] occurrences only; results for /ʀ/ remained stable. For /v/, the proportions of false rejects (FR) dropped from 39.1% to 12.0%, and CR precision doubled from 23.4% to 43.3%. It is coherent with the manual annotations in which the annotators often labeled as correct realizations of /v/ as [f], which is an acceptable pronunciation in given contexts. Thus, the classifier learned to be more permissive with these realizations, even with the ones that had a relatively high GOP score. The log-likelihood scores (*+2*) slightly decreased performance probably because they were redundant with GOP scores and noisy. The phonological feature (*+3*) brought improvement by increasing the CA rate from 50.8% to 52.0%. Correct realizations of /v/ as [f] benefited from this feature since /v/ and /f/ differ by a single distinctive feature: the voice feature. Adding the phone context identity (*+4*) brought useful information since SA increased 1.0% absolute. The phone and HMM middle-state duration ratio (*+5*) did not help, nor the CV ratio (*+6*), with which performance even dropped 14.3% absolute. This can be explained by the fact that vowel insertion is typical from Japanese speakers and these errors were not introduced in the training corpus. The best results (77.1% SA), shown in the last column of Figure 2, were achieved by concatenating the baseline features and the three features that brought improvement as single extra features: features 1, 3 and 4. As stated above, the annotator agreement was larger for /v/ than for /ʀ/ realizations. A similar trend was observed with the best system: accuracy for /v/ was much higher than the /ʀ/ one: 85.8% and 69.1%, respectively.

Finally, it is interesting to have a look at the LR weights of the best combination. The largest positive weights that favor the final decision towards the positive class (accept) involve "reco:R", "leftcontext:t", "reco:f", "reco:v" in decreasing order. The "reco:R" feature stands for the fact that the [ʀ] phone was recognized. It is indeed a positive feature when the expected target phone is [ʀ], and similarly with the "reco:v" feature for the [v] target phone. These features were expected to be important. The more surprising one is "reco:f", which means that the phone recognition system tends to recognize [f] instead of [ʀ] or [v] for occurrences that were judged as correct by the annotators. This illustrates a limit of the ASR-based approach due to the fact that the phone recognition is not always accurate. The second most positive feature was "leftcontext:t", which corresponds to the samples with a [tʀ] consonant cluster. It seems to indicate that words with this consonant cluster are not difficult to pronounce for the Japanese learners of our experiment. Finally, the largest negative weights favoring the mispronunciation decision involve the "reco:l" and "reco:v" features that correspond to the most frequent confusions made by Japanese learners for [ʀ] or [v], respectively.

5. Conclusions

In this paper, we reported pronunciation assessment experiments at phone-level of speech collected from Japanese learners of French as a foreign language. Our objective was to improve the accuracy of standard approaches, namely *Goodness-of-Pronunciation* and linear discriminant analysis on low-level acoustic features, as it is crucial for CAPT systems in order to be actually used by language learners. These baseline approaches were outperformed by the use of a logistic regression classifier on top of the F-GOP algorithm, thank to the possibility to add informative features as input to the classifier. A significant gain of 20.8% relative was obtained by adding phonetic and phonological features, leading to a 77.1% accuracy on a test corpus comprised of speech from 23 FFL Japanese speakers. To further improve these results, we plan to test model adaptation. Indeed, as the LR classifier was trained on a native speech corpus in which artificial errors were introduced, it may benefit from parameter adaptation with non-native speech material, even with little data. Another improvement direction involve testing more complex classifiers. Our recent experiments with convolutional neural networks with acoustic input features outperform LDA but not LR with the extra features introduced in the present study, so far. Finally, the manual annotations reflected that phone deviance greatly depends on intraword position. Phone position in words should then be taken into account when introducing artificial errors in the pronunciation lexicon.

6. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] B. Sevenster, G. d. Krom, and G. Bloothooft, "Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs," in *Proc. STiLL*, Marholmen, 1998, pp. 91–94.

[3] S. Witt, "Use of Speech Recognition in Computer-Assisted Language Learning," PhD Dissertation, University of Cambridge, Dept. of Engineering, 1999.

[4] H. Strik, K. P. Truong, F. de Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," in *Proc. INTERSPEECH*, 2007, pp. 1837–1840.

[5] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[6] Y. Tsubota, M. Dantsuji, and T. Kawahara, "An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors," *ReCALL*, vol. 16, pp. 173–188, 5 2004.

[7] S. Kanters, C. Cucchiarini, and H. Strik, "The Goodness of Pronunciation Algorithm: A Detailed Performance Study," in *SLaTE 2009 - 2009 ISCA Workshop on Speech and Language Technology in Education*, 2009, pp. 2–5.

[8] D. Le and E. M. Provost, "Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation." in *INTERSPEECH*, 2014, pp. 1563–1567.

[9] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation," in *Proc. Interspeech*, Brigthon, 2009, pp. 608–611.

[10] S. Witt and S. Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning," *Speech Communication*, vol. 30, pp. 95–108, 2000.

[11] J. Tomimoto and Y. Takaoka, "Le français, une langue imprononçcable pour les Japonais ?" *Rencontres Pédagogiques du Kansaï*, 2008.

[12] H. Yamasaki and P. Hallé, "How do native speakers of japanese discriminate and categorize french /r/ and /l/?" in *Proceedings of ICPhS*, San Francisco, 1999, pp. 909–912.

[13] S. Detey, J. Durand, and J.-L. Nespoulous, "Interphonologie et représentations orthographiques. Le cas des catégories /b/ et /v/ chez des apprenants japonais de Français Langue Etrangère," *Revue PArole*, vol. 34-35-36, pp. 140–185, 2005.

[14] S. Theodoridis, *Machine Learning*. Elsevier, 2015.

[15] J.-L. Gauvain, L. Lamel, and M. Eskenazi, "Design considerations and text selection for BREF, a large french read-speech corpus," in *Proc. ICSLP-90*, 1990, pp. 1097–2000.

[16] "PHON-IM project Web page," http://goo.gl/qwh709, [Online; accessed 20-September-2015].

[17] T. Kawahara and N. Minematsu, *Tutorial on CALL Systems at Interspeech*, Portland, 2012.

[18] S. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[19] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Proc. Interspeech*, 2005, pp. 1149–1152.

[20] G. Gravier, "ESTER train phase 1 IRISA phonetic alignments," http://goo.gl/ZVtKLY, 2005, [Online; accessed 20-September-2015].

[21] J. Farinas, "Multilingual phonetic decoders," http://www.irit.fr/recherches/SAMOVA/pagedap.html, 2013, [Online; accessed 20-September-2015].

[22] S. Detey, "Interphonologie et représentations orthographiques. Du rôle de l'écrit dans l'enseignement/apprentissage du français oral chez des étudiants japonais," Ph.D. dissertation, University of Toulouse-Le Mirail, 2005.

[23] T. Ooigawa, "Individual difference in production of voicing of french /r/ sounds and the perception by japanese adult listeners," in *Proceedings of CIL18*, 2009, pp. 1094–1111.

[24] L. Labrune, *The Phonology of Japanese*. Oxford: Oxford University Press, 2012.