IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Performance Comparison of Specific and General-Purpose ASR Systems for Pronunciation Assessment of Japanese Learners of Spanish

Cristian Tejedor-García[1], *Valentín Cardeñoso-Payo*[1], *David Escudero-Mancebo*[1]

[1]ECA-SIMM Research Group, Department of Computer Science, University of Valladolid, Spain
{cristian, valen, descuder}@infor.uva.es

Abstract

General–purpose state-of-the-art automatic speech recognition (ASR) systems have notably improved their quality in the last decade opening the possibility to be used in different practical applications, such as pronunciation assessment. However, the assessment of short words as minimal pairs in segmental approaches remains an important challenge for ASR, even more for non-native speakers. In this work, we use both our own tailored specific–purpose Kaldi–based ASR system and Google ASR to assess Spanish minimal pair words produced by 33 native Japanese speakers and to discuss their performance for computer-assisted pronunciation training (CAPT). Participants were split into three groups: experimental, in-classroom, and placebo. First two groups followed a pre/post-test training protocol spanning four weeks. Both the experimental and in-classroom groups achieved statistically significant differences at the end of the experiment, assessed by both ASR systems. We also found moderate correlation values between Google and Kaldi ASR systems in the pre/post-test values, and strong correlations between the post-test scores of both ASR systems and the CAPT application scores at the end of the experiment. Tailored ASR systems can bring clear benefits for a detailed study of pronunciation errors and results showed that they can be as useful as general–purpose ASR for assessing minimal pairs in CAPT tools.

Index Terms: automatic speech recognition (ASR), automatic assessment tools, foreign language pronunciation, pronunciation training, automatic pronunciation assessment, learning environments, minimal pairs

1. Introduction

Recent advances in automatic speech recognition (ASR) have made this technology a potential solution for transcribing audio input for computer-assisted pronunciation training (CAPT) tools [1, 2]. Available ASR technology, properly adapted, might help human instructors with pronunciation assessment tasks, freeing them from hours of tedious work, allowing for the simultaneous and fast assessment of several students, and providing a form of assessment that is not affected by subjectivity, emotion, fatigue, or accidental lack of concentration [3]. Thus, ASR systems can help in the assessment and feedback on learner productions, reducing human costs [4, 5]. Although most of the scarce empirical studies which include ASR technology in CAPT tools assess sentences in large portions of either reading or spontaneous speech [6, 7], the assessment of words in isolation remains a substantial challenge [8, 9].

This study has been partially supported by the Ministerio de Economía y Empresa (MINECO) and the European Regional Development Fund FEDER (TIN2014-59852-R) and by the Consejería de Educación of Junta de Castilla y León (VA050G18) and by the University of Valladolid (Ph.D. Research Grant 2015).

General–purpose off-the-shelf ASR systems like Google ASR[1] are becoming progressively popular each day due to their easy accessibility, scalability, and most importantly, effectiveness [10, 11]. These services provide accurate speech-to-text capabilities to companies and academics who might not have the possibility of training, developing, and maintaining a specific–purpose ASR system. However, despite the advantages of these systems (e.g., they are trained on large datasets and span different domains) there is an obvious need for improving their performance when used on in-domain data a specific scenarios, such as segmental approaches in CAPT for non-native speakers. Concerning the existing ASR toolkits, Kaldi has shown its leading role in recent years with its advantages of having flexible and modern code that is easy to understand, modify, and extend [12], becoming a highly matured development tool for almost any language [13, 14].

English is the most practiced language in CAPT experiments [6] and in commercial language learning applications, such as Duolingo[2] or Babbel[3]. However, there are scarce empirical experiments in the state-of-the-art which focus on pronunciation instruction and assessment for native Japanese learners of Spanish as foreign language, and as far as we are concerned, no one includes ASR technology. For instance, 1440 utterances of Japanese learners of Spanish as a foreign language (A1-A2) were analyzed manually with Praat by phonetics experts in [15]. Students performed different perception and production tasks with an instructor, and they achieved positive significant differences (at the segmental level) between the pre-test and post-test values. A pilot study on perception of Spanish stress by Japanese learners of Spanish was reported in [16]. Native and non-native participants listened to natural speech recorded by a native Spanish speaker and were asked to mark one of three possibilities (the same word with three stress variants) of an answer sheet. Non-native speech was manually transcribed with Praat by phonetic experts in [17], in an attempt to establish rule-based strategies for labeling intermediate realizations, helping to detect both canonical and erroneous realizations in a potential error detection system. Different perception tasks were carried out in [18]. It was reported how the speakers of native language (L1) Japanese tend to perceive Spanish /y/ when it is pronounced by native speakers of Spanish; and how the L1 Spanish and L1 Japanese listeners evaluate and accept various consonants as allophones of Spanish /y/, comparing both groups.

In previous work, we presented the development and the first pilot test of a CAPT application with ASR and text-to-speech technology, Japañol, through a training protocol [19, 20]. This learning application for smart devices includes a specific exposure–perception-production cycle of training activities

[1]https://cloud.google.com/speech-to-text
[2]https://www.duolingo.com/
[3]https://www.babbel.com/

10.21437/IberSPEECH.2021-2

with minimal pairs which are presented to students in lessons of the most difficult Spanish contrasts for native Japanese speakers. We were able to empirically measure statistically significant improvement between the pre and post-test values of 8 native Japanese speakers in a single experimental group. The students' utterances were assessed by experts in phonetics and by Google ASR system, obtaining strong correlations between human and machine values. After this first pilot test, we wanted to take a step further and to find pronunciation mistakes associated with key features of proficiency level characterization of more participants (33) and different groups (3). However, assessing such a quantity of utterances by human raters derived to a problem of time and resources. Also, Google ASR pricing policy and its limited black-box functionality also motivated us to look for alternatives to assess all the utterances, developing an own ASR system with Kaldi. In this work, we analyze the audio utterances of the pre-test and post-test of 33 Japanese learners of Spanish as foreign language with two different ASR systems (Google and Kaldi) to address the question of how these general and specific–purpose ASR systems can deal with the assessment of short words in the field of CAPT.

This paper is organized as follows. The experimental procedure is described in section 2, which includes the participants and protocol definition, a brief description about the process for elaborating the Kaldi–based ASR system, and the collection of metrics and instruments for collecting the necessary data. Results section shows the word error rate (WER) values of the Kaldi–based ASR system developed, the pronunciation assessment of the participants at the beginning and at the end of the experiment, including intra and inter-group differences, and the ASR scores' correlation of both ASR systems. We end this paper with a discussion about the performance of both state-of-the-art ASR systems in CAPT supported by our results and we shed light on lines of future work.

2. Experimental Procedure

2.1. Participants

A total of 33 native Japanese speakers from 18 to 26 years old participated voluntarily for the experimental prototype. All of them declared a low level of Spanish as foreign language with no previous training in Spanish phonetics. None of them stayed in any Spanish speaking country for more than 3 months. Besides, they were requested not to do any extra work in Spanish (e.g., conversation exchanges with natives or extra phonetics research) while the experiment was still active.

Participants were randomly divided into three groups: (1) **experimental group**, 18 students (15 female, 3 male) who trained their Spanish pronunciation with Japañol, during three sessions of 60 minutes; (2) **in-classroom group**, 8 female students who attended three 60-minutes pronunciation teaching sessions within the Spanish course, with their usual instructor, making no use of any computer-assisted interactive tools; and (3) **placebo group**, 7 female students who only took the pre-test and post-test. They did not attend neither the classroom nor the laboratory for Spanish phonetics instruction.

Finally, a group of 10 native Spanish speakers from the theater company Pie Izquierdo of Valladolid (5 women and 5 men) participated in the recording of a total of 41,000 utterances (7.1 hours of speech data) for the training corpus of the Kaldi ASR system for assessing the students' utterances gathered during the experimentation.

2.2. Protocol Description

We followed a four-week protocol which included a pre-test, three training sessions, and a post-test for the non-native participants. Native speakers recorded the speech training corpus for the Kaldi–based ASR system. At the beginning, the non-native subjects took part in the pre-test session individually in a quiet testing room. The utterances were recorded with a microphone and an audio recorder (the procedure was the same for the post-test). All the students took the pre-test under the sole supervision of a member of the research team. They were asked to read aloud the 28 minimal pairs administered via a sheet of paper with no time limitation[4]. The pairs came from 7 contrasts of the most difficult to perceive and produce Spanish consonant sounds by native Japanese speakers (see more details in [19]): [θ]–[f], [θ]–[s], [fu]–[xu], [l]–[ɾ], [l]–[r], [ɾ]–[rr], and [fl]–[fr]. Students were free to repeat each contrast as many times as they want if they thought they might have mispronounced them. Each participant took an average of 83.77 seconds to complete the pre-test (63.85 seconds min. and 129 seconds max.) and an average of 94.10 seconds to complete the post-test (52.45 and 138.87 seconds min. and max.).

From the same 7 contrasts, a total of 84 minimal pairs[4] were presented to the experimental and in-classroom group participants in 7 lessons along three training sessions. The minimal pairs were carefully selected by experts taking into account the Google ASR limitations (homophones, word-frequency, very short words, and out-of-context words, in a similar process as in [8]). The lessons were included in the CAPT tool for the experimental group and during the class sessions for the in-classroom group (12 minimal pairs per lesson, 2 lessons per session, except for the last session that included 3 lessons, see more details about the training activities in [19]). The training protocol sessions were carried out during students course's classes, in which a minimal pair was practiced in each lesson (blocked practice) and most phonemes were retaken in later sessions (spaced practice). Regarding the sounds practiced in each session, in the first one, sounds [fu]–[xu] and [l]–[ɾ] were contrasted. In the second one, [l]–[r] and [ɾ]–[rr]. The last session involved the sounds [fl]–[fr], [θ]–[f], and [θ]–[s]. Finally, subjects of the placebo group did not participate in the training sessions. They were supposed to take the pre-test and post-test and obtain results without significant differences. All participants were awarded with a diploma and a reward after completing all stages of the experiment.

On the other hand, each one of the native speakers recorded individually 164 words[4] for 25 times (41,000 utterances in total) presented randomly in five-hour sessions, for elaborating the training corpus for the Kaldi–based ASR system. The average, minimum, maximum, and standard deviation of the words length were: 4.29, 2, 8, and 1.07, respectively. The phoneme frequency (%) was: [a]: 16.9, [o]: 11.3, [r]: 9.0, [e]: 7.8, [f]: 5.3, [s]: 5.0, [ɾ]: 4.8, [l]: 4.5, [t]: 3.6, [k]: 3.6, [u]: 3.2, [i]: 3.2, [θ]: 3.2, [n]: 2.8, [m]: 2.3, [ɣ]: 1.8, [j]: 1.4, [ð]: 1.5, [x]: 1.3, [b]: 1.3, [p]: 1.1, [d]: 1.1, [β]: 0.9, [w]: 0.9, [ŋ]: 0.7, [g]: 0.3, [ʤ]: 0.2, and [z]: 0.1. The recording sessions were carried out in an anechoic chamber of the University of Valladolid with the help of a member of the ECA-SIMM research group.

2.3. Elaborating an ASR System with Kaldi

We analyzed the pre/post-test utterances of the participants with Kaldi and Google ASR systems. We did not have access to

[4] https://github.com/eca-simm/minimal-pairs-japanol-eses-jpjp

enough human resources to carry out the perceptual assessment of such a quantity of audio files, and Google ASR system just offered a limited black-box functionality and specification, so that, we developed our in–house Kaldi–based ASR system. In order to do so, different phoneme-level train models were tested in the Kaldi ASR system with the audio dataset recorded with native speakers before assessing the non-native test utterances.

The ASR pipeline that we have implemented uses a standard Gaussian Mixture Model-Hidden Markov Model (GMM/HMM) architecture, adapted from existing Kaldi recipes [12, 21]. After collecting and preparing the speech data for training and testing, the first step is to extract acoustic features from the audio utterances and training monophone models. To train a model, monophone GMMs are first iteratively trained and used to generate a basic alignment. Triphone GMMs are then trained to take surrounding phonetic context into account, in addition to clustering of triphones to combat sparsity. The triphone models are used to generate alignments, which are then used for learning acoustic feature transforms on a per-speaker basis in order to make them more suited to speakers in other datasets [22]. In our case, we re-aligned and retrained these models four times (tri4).

2.4. Instruments and Metrics

We gathered data from five different sources: (1) a registration form with student's demographic information, (2) pre-test utterances, (3) log files and (4) utterances of user's interaction with Japañol, and (5) post-test utterances. Personal information included name, age, gender, L1, academic level, and final consent to analyze all gathered data. Log files gathered all low-level interaction events with the CAPT tool and monitored all user activities with timestamps. From these files we computed a CAPT score per speaker which refers to the final performance at the end of the experiment. It includes the number of correct answers in both perception and production (in which we used Google ASR) tasks while training with Japañol [19]. Pre/post-test utterances consisted in oral productions of the minimal pairs lists provided to the students.

A set of experimental variables was computed: (1) WER values of the train/test set models for the specific–purpose Kaldi ASR system developed in a [0, 100] scale; (2) the student's pronunciation improvement at the segmental level comparing the difference of number of correct words at the beginning (pre-test) and at the end (post-test) of the experiment in a [0, 10] scale. We used this scale for helping teachers to understand the score as they use it in the course's exams. This value consists on the mean of correct productions in relation to the total of utterances. Finally, (3) the correlation values between Google and Kaldi ASR systems of the pre/post-test utterances and between the CAPT score and both ASR systems at the end of the experiment (post-test) in a [0, 1] scale.

By way of statistical metrics and indexes, Wilcoxon signed–rank tests have been used to compare the differences between the pre/post-test utterances of each group (intra-group), Mann–Whitney U tests have been used to compare the differences between the groups (inter-group), and Pearson correlations have been used to explain the statistical relationship between the values of the ASR systems and the final CAPT scores.

3. Results

Table 1 shows the models tested for native (Kaldi) and non-native (Kaldi and Google) speech data gathered, and the WER value reported by Google for natives [10]. Regarding the native models, the *All* model included 41,000 utterances of the native speakers in the train set. The *Female* model included 20,500 utterances of the 5 female native speakers in the train set. The *Male* model included 20,500 utterances of the 5 male native speakers in the train set. The *Best1*, *Best2*, and *Best3* models included 32,800 utterances (80%) of the total of native speakers (4 females and 4 males) in the train set. These last three models were obtained by comparing the WER values of all possible 80%/20% combinations (train/test sets) of the native speakers (e.g., 4 female and 4 male native speakers for training: 80%, and 1 female and 1 male for testing: 20%), and choosing the best three WER values (the lowest ones). On the other hand, the non-native test model consisted of 3,696 utterances (33 participants x 28 minimal pairs x 2 words per minimal pair x 2 tests).

Table 1: *WER values (%) of the ASR systems.*

	Train model						
		Kaldi					
	Google	*All*	*Female*	*Male*	*Best1*	*Best2*	*Best3*
Native	5.0	0.0024	3.10	1.55	0.14	0.14	0.23
Non-native	30.0	44.22	55.91	64.12	46.40	46.98	48.08

Google reported a 5.0% WER for their English ASR system for native speech [10]. Their training techniques are applied also for their ASR in other majority languages, such as Spanish. Thus, we extrapolated this WER value to our Spanish experiment. Regarding the Kaldi-based ASR system, we achieved values lower than 5.0% for native speech for the specific battery of minimal pairs introduced in Section 2 (e.g., *All* model: 0.0024%). On the other hand, we tested the non-native minimal pairs utterances with Google ASR obtaining a 30.00% (16% non-recognized words). In the case of the Kaldi–based ASR, as expected, the *All* model reported the best test results (44.22%) for the non-native speech. The *Female* train model derived into a better WER value for the non-native test model (55.91%) than the *Male* one (64.12%) since 30 out of 33 participants were female speakers.

Table 2 shows the mean scores assigned by the Google and Kaldi ASR systems to the 3,696 utterances of the pre/post-tests classified by the three groups of participants, in a [0, 10] scale. The students who trained with the tool (experimental group) achieved the best pronunciation improvement values in both Google (0.7) and Kaldi (1.1) ASR systems. However, the in-classroom group achieved better results in both tests and by both ASR systems (4.1 and 6.1 in the post-test; and 3.5 and 5.2 in the pre-test, Google and Kaldi, respectively). The placebo group achieved the worst post-test (3.2 and 3.5, Google and Kaldi, respectively) and pronunciation improvement values (0.2 and 0.4, Google and Kaldi, respectively).

A Wilcoxon signed–rank test found statistically significant intra-group differences between the pre- and post-test values of the experimental and in-classroom groups of both ASR systems. In the case of the placebo group, there were differences only in the Google ASR values (see p and Z values in Table 2). Concerning inter-group pairs comparisons, a Mann–Whitney U test found statistically significant differences between the experimental and in-classroom groups in the post-test Google ASR scores ($p < 0.001$; Z = -2.773) and Kaldi ones ($p < 0.001$; Z = -2.886). There were also differences between the experimental and placebo groups in the post-test Kaldi scores ($p < 0.001$;

Table 2: *Pre/post-test scores. $\overline{n}$, N, and Δ refer to mean score of the correct pre/post-test utterances, number of utterances, and difference between the post and pre-test mean scores, respectively.*

	Pre-test				Post-test				Δ (Post-test - Pre-test) – Wilcoxon signed–rank test					
	Google		Kaldi		Google		Kaldi		Google			Kaldi		
Group	$\overline{n}$	N	$\overline{n}$	N	$\overline{n}$	N	$\overline{n}$	N	Δ	p-value	Z	Δ	p-value	Z
Experimental	3.0	560	4.1	560	3.7	560	5.2	560	0.7	< 0.001	-13.784	1.1	< 0.001	-5.448
In-classroom	3.5	448	5.2	448	4.1	448	6.1	448	0.6	< 0.001	-2.888	0.9	< 0.001	-3.992
Placebo	3.0	392	3.1	392	3.2	392	3.5	392	0.2	0.002	-3.154	0.4	0.059	-1.891

Z = -5.324). Post-test differences between the in-classroom and placebo groups were only found in the Kaldi scores ($p < 0.001$; Z = -7.651). Finally, although there were significant differences between the pre-test scores of the in-classroom group and the experimental group (Google: $p < 0.001$; Z = -8.892; Kaldi: $p < 0.001$; Z = -3.645), and the placebo group (Google: $p < 0.001$; Z = -8.050; Kaldi: $p = 0.001$; Z = -3.431), such differences were minimal since the effect size values were small (r = 0.10 and r = 0.20, respectively).

Table 3: *Regression coefficients of the ASR and CAPT systems. x, y, a, b, $S.E.$, and r refer to dependent variable, independent variable, slope of the line, intercept of the line, standard error, and Pearson coefficient, respectively.*

x	y	a	b	$S.E.$	r	p-value
pre-Kaldi	pre-Google	0.927	1.919	0.333	0.51	0.005
post-Kaldi	post-Google	0.934	1.897	0.283	0.57	0.002
post-Google	CAPT	0.575	-0.553	0.148	0.81	0.002
post-Kaldi	CAPT	0.982	-1.713	0.314	0.74	0.007

Finally, we analyzed several correlations between (1) the pre/post-test scores of both ASR systems (three groups) and (2) the CAPT scores with the experimental group post-test scores of both ASR systems (only group with a CAPT score) in order to compare the three sources of objective scoring (Table 3). The first and second rows of Table 3 represent the moderate positive Pearson correlations found between the Google and Kaldi pre-test (r = 0.51, p = 0.005) and post-test (r = 0.57, p = 0.002) scores. Finally, the third and fourth rows of Table 3 represent the fairly strong positive Pearson correlations found between the CAPT scores and the post-test scores of Google (r = 0.81, p = 0.002) and Kaldi (r = 0.74, p = 0.007) ASR systems.

4. Discussion and Conclusions

We have reported on empirical evidences about significant pronunciation improvement at the segmental level of native Japanese beginner–level speakers of Spanish by using state-of-the-art ASR systems (Table 2). In particular, the experimental and in-classroom group speakers improved 0.7|1.1 and 0.6|0.9 points out of 10, assessed by Google|Kaldi ASR systems, respectively, after just three one-hour training sessions. These results agreed with those reported in [8, 23]. Thus, the training protocol and the technology included, such as the CAPT tool and the ASR systems provided a very useful and didactic instrument that can be used complementary with other forms of second language acquisition in larger and more ambitious language learning projects.

Our specific–purpose Kaldi ASR system allowed us to reliably measure the pronunciation quality of the substantial quantity of utterances recorded. In particular, this ASR system proved to be useful for working at the segmental (phone) level for non-native speakers. Developing an in–house ASR system allowed us not only to customize the post-analysis of the speech without the black-box and pricing limitations of the general–purpose Google ASR system, but also neither pre-discard specific words (e.g., infrequent, out-of-context, and very short words) nor worry about the data privacy. Despite the positive results reported about the Kaldi ASR, the training corpus was limited in both quantity and variety of words and the experiment was carried out under a controlled environment. Noise-reduction, data augmentation, and a systematic study of the non-native speech data gathered to find pronunciation mistakes associated with key features of proficiency level characterization with the help of experts for its automatic characterization [4, 17] must be considered in the future to expand the project.

We have also compared our Kaldi ASR results with Google ASR ones, obtaining moderate correlations between them (Table 3). Although our specific–purpose ASR system is neither as accurate nor ambitious as Google ASR, it seems to be promising and robust enough for a concrete battery of minimal pairs. Hence, both state-of-the-art ASR systems proved to be valid for our pronunciation assessment task of minimal pair words. Future work will consist on a fine–tuning of our Kaldi–based ASR system with more utterances and re-training techniques, such as deep or recurrent neural networks, combining both native and non-native speech in order to improve current results and to obtain a better customization of the ASR system to the specific phone–level tasks. Thus, researchers, scholars, and developers can decide which one to integrate in their CAPT tools depending on the tasks and resources available.

Finally, the post-test values of both Google and Kaldi ASR systems strongly correlated with the final scores provided by the CAPT tool of the experimental group speakers (Table 3). That is, although the training words in Japañol were not the same as the pre/post-test ones, the phonemes trained were actually the same and the speakers were able to assimilate the lessons learned from the training sessions to the final post-test. Therefore, we were able to ensure that both scoring alternatives are valid and can be used for assessing Spanish minimal pairs for certain phonemes and contexts (e.g., resources availability, learning, place, data privacy, or costs).

5. Acknowledgements

The authors would like to thank Mr. Takuya Kimura for his support, the participants of the University of Seisen (Japan) and Language Learning Center of University of Valladolid, and the theater company Pie Izquierdo of Valladolid.

6. References

[1] E. Martín-Monje, I. Elorza, and B. G. Riaza, *Technology-Enhanced Language Learning for Specialized Domains: Practical Applications and Mobility*. Oxon, UK: Routledge, 2016. [Online]. Available: https://doi.org/10.4324/9781315651729

[2] O'Brien *et al.*, "Directions for the future of technology in pronunciation research and teaching," *J. Second Lang. Pronunciation*, vol. 4, no. 2, pp. 182–207, Feb. 2018. [Online]. Available: https://doi.org/10.1075/jslp.17001.obr

[3] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer-assisted pronunciation training," *Comput. Assisted Lang. Learn.*, vol. 15, no. 5, pp. 441–467, Aug. 2010. [Online]. Available: https://doi.org/10.1076/call.15.5.441.13473

[4] J. v. Doremalen, C. Cucchiarini, and H. Strik, "Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch," *J. Acoustical Soc. America*, vol. 134, no. 2, pp. 1336–1347, 2013. [Online]. Available: https://doi.org/10.1121/1.4813304

[5] T. Lee *et al.*, "Automatic speech recognition for acoustical analysis and assessment of Cantonese pathological voice and speech," in *Proc. ICASSP*, Shanghai, China, Mar. 20-25, 2016, pp. 6475–6479. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7472924

[6] R. I. Thomson and T. M. Derwing, "The effectiveness of L2 pronunciation instruction: A narrative review," *Appl. Linguistics*, vol. 36, no. 3, pp. 326–344, Jul. 2015. [Online]. Available: https://doi.org/10.1093/applin/amu076

[7] G. Seed and J. Xu, "Integrating technology with language assessment: Automated speaking assessment," in *Proc. ALTE*, Bologna, Italy, May 3–5, 2017, pp. 286–291.

[8] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool," *IEEE Trans. Learn. Technol.*, vol. 13, no. 2, pp. 269–282, Mar. 2020. [Online]. Available: https://doi.org/10.1109/TLT.2020.2980261

[9] J. Cheng, "Real-time scoring of an oral reading assessment on mobile devices," in *Proc. Interspeech*, Hyderabad, India, Sep. 2–6, 2018, pp. 1621–1625. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-34

[10] M. Meeker, "Internet trends 2017," may 2017, Kleiner Perkins, Los Angeles, CA, USA, Rep. [Online]. Available: https://www.bondcap.com/report/it17.

[11] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx)," *Int. J. Eng. Res. Appl*, vol. 7, no. 03, pp. 20–24, 2017. [Online]. Available: https://doi.org/10.9790/9622-0703022024

[12] D. Povey *et. al*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Waikoloa, Hawaii, HI, USA, Dec. 11–15, 2011, pp. 1–4.

[13] P. Upadhyaya, S. K. Mittal, O. Farooq, Y. V. Varshney, and M. R. Abidi, "Continuous Hindi Speech Recognition Using Kaldi ASR Based on Deep Neural Network," in *Mach. Intell. Signal Anal.*, M. Tanveer and R. B. Pachori, Eds. Singapore: Springer Singapore, 2019, pp. 303–311. [Online]. Available: https://doi.org/10.1007/978-981-13-0923-6_26

[14] I. Kipyatkova and A. Karpov, "DNN-Based Acoustic Modeling for Russian Speech Recognition Using Kaldi," in *Speech Comput.*, A. Ronzhin, R. Potapova, and G. Németh, Eds. Cham: Springer International Publishing, 2016, pp. 246–253. [Online]. Available: https://doi.org/10.1007/978-3-319-43958-7_29

[15] G. F. Lázaro, M. F. Alonso, and K. Takuya, "Corrección de errores de pronunciación para estudiantes japoneses de español como lengua extranjera," *Cuadernos CANELA*, vol. 27, pp. 65–86, Jan. 2016.

[16] T. Kimura, H. Sensui, M. Takasawa, A. Toyomaru, and J. J. Atria, "A Pilot Study on Perception of Spanish Stress by Japanese Learners of Spanish," in *Proc. SLATE*, Tokyo, Japan, Sep. 22–24, 2010.

[17] M. Carranza, "Intermediate phonetic realizations in a Japanese accented L2 Spanish corpus," in *Proc. SLaTE*, Grenoble, France, Aug./Sep. 30–1, 2013, pp. 168–171.

[18] T. Kimura and T. Arai, "Categorical Perception of Spanish /y/ by Native Speakers of Japanese and Subjective Evaluation of Various Realizations of /y/ by Native Speakers of Spanish," *Speech Res.*, vol. 23, pp. 119–129, 2019.

[19] C. Tejedor-García, V. Cardeñoso-Payo, M. J. Machuca, D. Escudero-Mancebo, A. Ríos, and T. Kimura, "Improving Pronunciation of Spanish as a Foreign Language for L1 Japanese Speakers with Japañol CAPT Tool," in *Proc. IberSPEECH*, Barcelona, Spain, Nov. 21–23, 2018, pp. 97–101. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-21

[20] C. Tejedor-García, V. Cardeñoso-Payo, and D. Escudero-Mancebo, "Japañol: a mobile application to help improving Spanish pronunciation by Japanese native speakers," in *Proc. IberSPEECH*, Barcelona, Spain, Nov. 2018, pp. 157–158.

[21] E. Chodroff, "Corpus Phonetics Tutorial," 2018. [Online]. Available: https://arxiv.org/abs/1811.05553

[22] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance Gaussians," in *Proc. Interspeech*, Pittsburgh, PA, USA, Sep. 17–21, 2006, pp. 1145–1148.

[23] C. Tejedor-García, D. Escudero-Mancebo, V. Cardeñoso-Payo, and C. González-Ferreras, "Using challenges to enhance a learning game for pronunciation training of English as a second language," *IEEE Access*, vol. 8, no. 1, pp. 74 250–74 266, Apr. 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2988406