9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Analysis of the efficiency of repeating activities for improving prosody in L2 pronunciation training

David Escudero-Mancebo, César González-Ferreras, Valentín Cardeñoso-Payo

ECA-SIMM Research Group
University of Valladolid, Spain
descuder@infor.uva.es

Abstract

Repeating activities are frequently proposed in courses and tools for improving foreign language pronunciation. In this work we present a study that aims to quantify experimentally the degree of improvement that students reach by performing such activities in what concerns to prosody. A group of Japanese and American students of L2 Spanish read several times a set of sentences in different conditions (listening-and-reading or only reading). Subjective scoring of the utterances was performed by following a set of quality criteria. Additionally the objective scoring at suprasegmental level of the utterances was also measured with a set of objective metrics that have to do with temporal, energy and fundamental frequency domains. Results prove that foreign utterances are closer to the reference ones after repetitions and fluency increases both from subjective and objective scores. It is not clear that other particular problems such as accent and rhythm also improve without specific feedback.

Index Terms: computer assisted pronunciation training, comparing prosody[1]

1. Introduction

Pronunciation is one of the main language learning dimensions. Computer supported systems and speech technologies have entered into language learning domain leading to the creation of the disciplines of Computer Assisted Language Learning (CALL) and Computer Assisted Pronunciation Training (CAPT). In [1] we reviewed the main activities that are present in nowadays CAPT systems for pronunciation training. Among them, reading and repeating are frequently found. For example in modern language learning web services, it is common to find activities in which the system proposes a word or a sentence to be read by the user for the system to rate the quality of the utterance; the user must repeat the utterance until its quality is good enough. This reading can be supported by the listening of a correct pronunciation of the target text, in which case the activity could be named *parroting*. Imitation drills and reading aloud are still the most common elocution exercises in teaching pronunciation [2]. Most language pronunciation teaching approaches at segmental level in CALL start from the idea that perceiving sounds which do not exist in L1 is an essential prerequisite for good pronunciation in L2 ([3]). Recent results suggest a positive impact of imitation computer-based phonetic training on L2 sound perceptual awareness [4]. Imitation of native speech templates through parroting exercises could make speech more native-like and has become a common practice in many mobile based CAPT tools. Combinations of suprasegmental and pronunciation scores have been successfully applied to the automatic assessment of nonnative pronunciation of L2 [5]. The effects of oral repetition and practice in improving fluency in L2 acquisition have recently been reported [6]. In this paper we present an evaluation of the impact on pronunciation quality of repeating the readings in working sessions, in particular the impact on prosodic quality.

Evaluation in conventional language learning courses is responsibility of teachers who, after listening to the students, and at the glance of previously established learning goals, judge the students competences. Recently, there are alternatives to this model, in which an electronic service scores the quality of the users pronunciation by analyzing speech with automatic methods (see table 2 of [7]). In this work, we evaluate the impact of repetitions by using both human judgments and automatic measurements that are obtained from acoustic correlates of prosody belonging to energy, F0 and duration domains.

The paper is oriented to answer the following research questions:

RQ1 Does the quality of prosody improve after repeating the reading exercises?

- **Issue 1.1** Improvements that are identified by human evaluators, can also be identified by automatic measurements?
- **Issue 1.2** Which particular aspects of prosody improve the most?

RQ2 What is the impact of parroting? Is it better that students listen to the correct pronunciation before reading?

The paper is organized as follows: first experimental procedure is detailed by presenting the corpus, the human evaluation and the automatic metrics to be used; next results section presents the improvements along repetitions both by using human and automatic rates; paper ends with discussion, conclusions and future work.

2. Experimental procedure

2.1. Corpus description

The corpus used in this work is described in detail in [8]. We recorded 14 Spanish L2 speakers: 9 American English and 5 Japanese. All of them were students of Spanish at a university level. We also recorded 8 native Spanish speakers of different speaking styles, to have a set of reference pronunciations. The set of foreign speakers was selected with the guidance of educational personnel of the Languages Center of our University, among students ranging from A2 to B2 Spanish proficiency levels.

[1]We would like to thank Ministerio de Economía y Competitividad y Fondos FEDER project key: TIN2014-59852-R Videojuegos Sociales para la Asistencia y Mejora de la Pronunciación de la Lengua Española

10.21437/SpeechProsody.2018-61

For every foreign speaker, each recording session included first sight read sentences, listen and repeat sentences, short stories and news paragraphs reading. For this work we select the first two blocks of reading activities which are described as follows:

- *First sight read sentences*. Fifteen short sentences were selected from the news paragraphs of the prosodic GLISSANDO corpus, following a phonetic coverage criterion (sentences are presented in cite [8]). From them, 10 (s01-s10) were selected to be read at first sight by non-native speakers. Ten sentences were read with small pauses between them and the task was repeated three times with resting stops in between. This provides a basis for the experimental study of the influence of simple reading repetition on the pronunciation correctness.
- *Listen and repeat sentences*. A group of 10 (s05-s15) additional sentences was gathered reusing the last 5 of the previous ten sentences and 5 fresh ones from the original set of fifteen sentences. Using a simple tablet application, a reference utterance of each sentence by a native professional speaker was presented to the non-native speaker, who had to carefully listen and repeat it immediately afterwards. Again, this process was repeated three times to provide a means of evaluation of the effectiveness of this guided pronunciation scheme.

2.2. Subjective evaluation of prosodic quality

Four experts have independently assigned **perceptual evaluation measures** along five different dimensions, using a Likert scale, and a proposed overall proficiency level according to the Common European Framework of Reference for Languages, Teaching and Assessment (CEFR) as applied to Spanish (DELE[2]).

All the labelers already had good competences in the evaluation of Spanish as a second language, developed as part of their training background in the university degree in Spanish Language and Literature. After a selection process, we provided specific training sessions on the evaluation protocol and the expected meaning and scales of the target parameters we proposed to label the utterances in the corpus. Open discussions favored the establishment of a common ground for the criteria to follow for the evaluation along the different dimensions.

The labeling process was monitored in order to detect possible anomalous deviations in the assessment criteria for some of the evaluators. Along the labeling process, we conducted several follow up sessions to try to keep general criteria as homogeneous as possible.

Most of the previous works have used a single dimension to assess pronunciation quality by human experts [9, 10, 11, 12, 13]. In this work, we follow an approach based on several dimensions, similar to the one recently proposed in [14], because this allows us to evaluate different aspects of the utterances instead of a single overall performance. Perceptual dimensions include:

- *intelligibility* (*eint*): the expert provides an integer value to indicate the level of understanding of what has been said (1:very poor, 5:excellent).
- *fluency* (*eflu*): the expert provides an integer value to indicate the level of interruptions, hesitations, filled pauses and other phenomena which could affect fluency (1:very poor, 5:excellent).
- *phonetic correctness* (*efon*): the expert provides an integer value in order to evaluate if all the phonemes have been correctly pronounced (1:clearly non-native, 5:native).
- *lexical accent correctness* (*eacc*): the expert provides an integer value in order to evaluate if lexical accent (position of the accented syllable within the word) is correctly positioned according to any accepted pronunciation of Spanish (1:clearly non-native, 5:native).
- *rhythm* (*erit*): the expert provides an integer value in order to evaluate to which extent the prosody clearly resembles the one in a native Spanish speaker or, on the contrary, shows a neat non-native accent (1:clearly non-native, 5:native).
- *Spanish level* (*edele*): the expert indicates which level of proficiency of Spanish appears to have the speaker, according to the DELE scheme (A1, A2, B1, B2, C1 or C2) and using a 1 (A1) to 6 (C2) numeric scale.

The labelers filled their evaluation scores for the perception experiment using a web-based application. A total of 1179 utterances were randomly presented to the evaluator in sequence through a web form. They could listen to the utterance as many times as they wanted and the form was filled with the perceptual scores, the estimated DELE reference level and any additional comments they would like to add for that particular utterance or speaker. The average evaluation time was around 9 times longer than the average utterance duration, which illustrates the high cost of manual annotation. Since the samples were presented at random, the likelihood that the labeler could listen to two of them in the same order they were recorded is negligible, as can be easily computed.

2.3. Objective evaluation of prosodic quality

The following metrics have been computed per sentence:

2.3.1. Duration related measurements

Speech rate measures: For each utterance, we compute a rate of speech (*ros*) as the number of phones per second.

Global interval proportions: We computed the vocalic intervals ratio (*VIR*) (sum of the lengths of vocalic intervals divided by the total duration of the sentence, excluding pauses), as proposed by [15]. The standard deviation of the duration of vocalic intervals (*dV*) and of consonantal intervals (*dC*) are computed at utterance level. Following [16], we also computed the standard deviation of consonantal (*varV*) and vocalic (*varC*) interval durations divided by mean consonantal or vocalic duration within the utterance.

Variability indexes: We identify vocalic and consonantal segments and computed two forms of the Pairwise Variability Index proposed in [17]:

$$rPVI = 100 \times \frac{\sum_{i=1}^{N-1} |d_i - d_{i+1}|}{N-1} \quad (1)$$

$$nPVI = 100 \times \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|d_i - d_{i+1}|}{(d_i + d_{i+1})/2} \quad (2)$$

With these, four utterance-level features are extracted: *rPVI.V, nPVI.C, rPVI.V, rPVI.C* separating consonant and vocalic segments.

[2]http://www.dele.org/

2.3.2. Energy and F0 related measurements

First, we obtained energy and F0 of each utterance at intervals of 10 milliseconds using praat [18]. Energy was normalized by speaker and F0 was measured in semitones with respect to the mean F0 of each speaker. To compare two utterances we first aligned them using the energy and the dynamic time warping (DTW) algorithm. We obtained de following metrics:

DTW similarity indexes: *DTW.En* is the similarity measure provided by the DTW algorithm, divided by the sum of the lengths of the two input utterances. *DTW.F0* is calculated by using the DTW algorithm dividing the similarity index by the sum of the lengths of the two input utterances. We used the implementation of the DTW algorithm described in [19].

Once we have both utterances aligned with DTW (using energy), we computed the measures related to F0: the root-mean-square error (RMSE), the correlation which have been conventionally used for comparing F0 contour in speech synthesis [20].

RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (3)$$

where n is the number of the number of F0 values in the aligned utterances; X_i refers to the F0 values of the reference utterance and Y_i refers to the F0 values of the other utterance.

As F0 is not defined in unvoiced regions, different calculations of RMSE can be done. *RMSE* was calculated using only the values in which both utterances are voiced.

Pearson correlation is calculated as follows:

$$Cor = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (4)$$

Giving to X and Y the same meaning as before. Cor is calculated using only the values in which both utterances are voiced.

3. Results

Table 1 shows differences between values of the metrics described in previous section for native and non-native speakers. The values obtained by the native speakers are to be a reference, so that the closer to non-native speakers the values are, the best their pronunciation can be considered to be. Concerning the subjective metrics results, as expected, native speakers utterances obtained the maximum rate (or very close to it) at the time that non-native speakers obtain lower results. Concerning to the objective metrics, native speakers obtain higher values for duration metrics *ROS*, *VIR* and for the F0 related metric *cor*. On the other hand, native speakers get lower results for the *d*, *var* and *PVI* duration metrics and for the RMSE, and DTW related metric. In the discussion section we are presenting an interpretation for these values, but now, we are interested in using these differences as a reference for assessing the impact of the repetitions in the non-native pronunciation quality.

Table 3 compares the marks assigned by the different evaluators to the different sentence repetitions. In the cells, the mean difference and its statistical significance is presented. The differences are computed as the difference between the marks obtained in the third and first repetition of the same sentence uttered by the same speaker; thus we average per sentence and per speaker. A positive value in row *R01 vs R03* means the marks assigned to the third repetition are higher (in average) than the marks assigned to the first one. In general terms, most of the values are positive, but the most clear differences appear in what respects to fluency. Repetition without listening activities (R01 vs R03 rows) permits to improve the marks with significant results in what concerns to *fluency* and overall *level* perception (except for evaluator B). A and B raters appreciate differences on *rhythm* and A and D marked differences on *intelligibility*. In parroting activities, A, C, D raters observed differences on *fluency* at the time that B indicated differences affecting *rhythm*. When reading and parroting are compared in row *R01 vs P01*, all the evaluators observed significant improvements on *fluency* but there are not significant improvements affecting the other indicators: only the evaluator B observed an improvement in overall *level*.

Table 1: *Difference between native and non-native speakers. min and max are the boundaries of the 95% confidence interval of the values of the variables per group of speakers.*

	Native		no-Native	
Metric	min	max	min	max
eint	5.00	5.00	3.40	3.46
eflu	4.92	4.99	3.20	3.27
efon	4.91	4.98	2.85	2.91
eacc	4.93	4.99	3.21	3.27
erit	4.79	4.91	2.53	2.59
edele	5.97	6.00	3.32	3.39
ROS	11.68	12.34	9.60	9.81
VIR	47.43	49.35	44.87	46.02
dV	0.04	0.05	0.07	0.08
dC	0.04	0.04	0.09	0.09
varV	55.30	62.44	73.36	76.57
varC	43.42	46.19	75.73	79.52
rPVI.V	3.70	4.19	6.72	7.17
rPVI.C	4.05	4.39	8.22	8.78
nPVI.V	42.57	45.41	56.72	58.52
nPVI.C	49.42	52.24	62.35	64.05
DTW.En	0.10	0.12	0.16	0.16
RMSE	2.86	3.37	3.34	3.42
Cor	0.51	0.61	0.26	0.29
DTW.F0	0.50	0.60	0.68	0.71

Tables 2 and 4 shows the objective metrics differences along the different activities. Positive values are obtained for *ROS* and *Cor* and negative values for the rest of metrics. This result is consistent with the expectations in row *R01 vs R03*, as it indicates that the prosodic production is closer to the reference after repeating (higher Cor, lower $DTW.En$, $RMSE$, dV, dC, $varC$ and PVI metrics) and faster (higher ROS). In the parroting activities (*P01 vs P03* row) significant changes are observed in *ROS*, dV, dC and PVI metrics. When reading and parroting are compared in row *R01 vs P01*, the highest improvements of the three tests are obtained for ROS, dV, dC, $varC$, PVI and Cor.

Table 2: *Differences of the objective metrics related with duration values among repetitions. The rows legend and cell values have the same meaning as in table 3*

Test	ROS	VIR	dV	dC	varV	varC	rPVI.V	rPVI.C	nPVI.V	nPVI.C
R01 vs R03	0.75 ****	-0.24 ns	-0.01 ***	-0.01 ***	-1.41 ns	-4.00 ***	-0.75 ****	-0.78 ****	-2.38 **	-3.44 ***
P01 vs P03	0.36 ****	-0.05 ns	-0.00 *	-0.00 *	-2.08 ns	-0.97 ns	-0.32 *	-0.38 **	0.05 ns	-1.19 ns
R01 vs P01	1.18 ****	-1.14 ns	-0.01 ****	-0.01 ***	-1.01 ns	-5.36 *	-1.49 ****	-1.35 ***	-6.33 ****	-4.34 *

Table 3: *Differences of the human ratings among repetitions. Eval is the human evaluator. R01 vs. R03 compares the ratings of the first and third repetitions of the utterance in reading without listening activities. P01 vs. P03 refers to the first and third repetitions of the utterance in parroting activities. R01 vs. P01 compares the ratings of the first reading of the sentence without and with parroting support respectively. ns means p> 0.05, * is p≤0.05, ** is p≤0.01, *** is p≤0.001, **** is p≤0.0001 when the paired t-test is applied to the samples of corpus.*

Eval	Test	Int	Flu	Pho	Acc	Ryt	Lev
A	R01 vs R03	0.24 *	0.34 **	0.05 ns	0.15 ns	0.20 ***	0.25 **
	P01 vs P03	0.17 ns	0.29 **	0.01 ns	0.16 ns	0.07 ns	0.27 **
	R01 vs P01	-0.13 ns	0.62 **	-0.13 ns	0.16 ns	0.24 ns	0.04 ns
B	R01 vs R03	0.23 ns	0.53 ***	0.05 ns	0.23 ns	0.23 *	0.28 ns
	P01 vs P03	0.12 ns	0.29 ns	0.17 ns	0.20 ns	0.22 *	0.27 ns
	R01 vs P01	0.00 ns	0.71 *	0.24 ns	0.33 ns	0.29 ns	0.48 *
C	R01 vs R03	0.05 ns	0.31 ****	0.08 ns	0.11 *	0.10 ns	0.25 ***
	P01 vs P03	0.14 ns	0.21 **	0.06 ns	0.07 ns	0.05 ns	0.14 *
	R01 vs P01	-0.10 ns	0.30 **	0.00 ns	0.04 ns	0.10 ns	0.09 ns
D	R01 vs R03	0.14 *	0.34 ***	0.06 ns	0.07 ns	0.05 ns	0.21 **
	P01 vs P03	0.05 ns	0.25 **	0.01 ns	-0.01 ns	0.04 ns	0.11 ns
	R01 vs P01	-0.06 ns	0.43 **	0.06 ns	-0.07 ns	0.06 ns	0.07 ns

Table 4: *Differences of the objective metrics related with F0 and energy values among repetitions. The rows legend and cell values have the same meaning as in table 3*

	DTW.En	RMSE	Cor	DTW.F0
r01 vs r03	-0.01 ***	-0.08 *	0.04 *	0.01 ns
p01 vs p03	0.00 ns	-0.06 ns	0.03 ns	0.02 ns
r01 vs p01	-0.01 ns	-0.09 ns	0.09 **	-0.04 ns

4. Discussion

Repeating the reading of isolated sentences permits to improve the prosodic quality of its production, mainly in what concerns to oral fluency (column Flu of row $R01$ vs $R03$ in table 3). The observed improvement on fluency is important as it contributes to improve the general perception of quality (column $Level$ of tabla 3). Apart from fluency, other prosody related aspects such as rhythm or accent do not improve in the same degree, probably because they need some feedback that is absent in the repeating exercises.

Listening-and-reading ($parroting$), seems to be an important help that contributes to obtaining improvements when utterances are compared with only reading activities: significant differences in row $R01$ vs $P01$. Again, fluency is the aspect which improves the most but it doesn't seem to be enough for improving rhythm, phonetic quality and accent. Repetition of $parroting$ activities seems to be less efficient than repetition of reading alone activities (no significant differences in row $P01$ vs $P03$).

The computed objective metrics show important differences between non-native and native prosody. They also show improvements along the repetitions which permits to defend its use in computer assisted pronunciation training. Concerning metrics related with duration, rate of speech is the metric that changes the most with repetitions, which can be explained because of its relation with fluency. Nevertheless other metrics like PVI, that have been traditionally related with rhythm, also change considerably.

The impact of listening before reading is clearly observed when the F0 distance is observed (row $R01$ vs $P01$). The important improvement affecting Cor (0.09 normalized points over starting values in the range of 0.26, 0.29 is an increase of 30%) is a clear indicator of the efficiency of parroting, as speakers read the target sentence by using a closer F0 to the reference one in $P01$ than in $R01$.

Human evaluators rate the utterances at the level of sentences. This fact make it difficult analyzing problems that occur at a lower linguistic level such as phrases, words or syllables leading to important inconsistencies among judgments (already reported in [8]). Objective metrics could contribute to detect differences that occur at these minor linguistic levels. In [21] we showed that automatic prosodic labels and the definition of specific metrics also permits to separate native and non-native utterances and evaluating the quality along repetitions. It is our current work using prosodic labels for identifying words or groups of words in the whole sentence that are wrongly pronounced.

As a limitation of the study, we must say that the investigation is focused on repetitions done in the same working session. Thus, very little can be said about the generalization of the improvements, which testing would require of the analysis of medium or long term working sessions.

5. Conclusions

In this paper we have presented and experiment for evaluating the goodness of repetitions as an efficient improving pronunciation exercise. Both repeating the reading of sentences and listening-and-reading activities permit to improve fluency. A set of objective metrics show to be efficient for detecting such improvements.

At the time that we have shown the efficiency of repetitions for improving pronunciation and the usefulness of the objective metrics, results also evidence that improvement affects mainly to fluency and that for improving other prosodic characteristics such as rhythm or accent position, repetition activities should, probably, be complemented with exercises that include some kind of feedback.

6. References

[1] D. Escudero-Mancebo and M. Carranza, "Nuevas propuestas tecnológicas para la práctica y evaluación de la pronunciación del español como lengua extranjera," in *Actas del L Congreso de la Asociación Europea de Profesores de Español, Burgos*, 2015, pp. 218–227.

[2] R. H. Jones, "Beyond listen and repeat : pronunciation teaching materials and theories of second language acquisition," *System*, vol. 25, no. 1, pp. 103–112, 1997.

[3] M. G. Busa, "New perspectives in teaching pronunciation," 01 2008.

[4] E. G. Lacabex and F. Gallardo del Puerto, "Two phonetic-training procedures for young learners: Investigating instructional effects on perceptual awareness," *Canadian Modern Language Review*, vol. 70, no. 4, pp. 500–531, 2014.

[5] J. Tepperman, T. Stanley, K. Haciglou, and B. Pellom, "Testing suprasegmental English through parroting," in *Speech Prosody*, 2010.

[6] Y. Yoshimura and B. Macwhinney, "The effect of oral repetition on L2 speech fluency: An experimental tool and language tutor," in *Workshop on Speech and Language Technology in Education, SLaTE 2007, Farmington, PA, USA, October 1-3, 2007*. Carnegie Mellon University / ISCA, 2007.

[7] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *International Symposium on Automatic Detection of Errors in Pronunciation Training, Stockholm, Sweden*, 2012.

[8] D. Escudero-Mancebo, C. González-Ferreras, and V. Cardeñoso Payo, "Assessment of non-native spoken spanish using quantitative scores and perceptual evaluation," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3967–3972.

[9] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and M. K. Snmez, "Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners." in *INTERSPEECH 2000*. ISCA, 2000, pp. 187–190.

[10] Y. Yamashita, K. Kato, and K. Nozawa, "Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison." *IEICE Transactions*, vol. 88-D, no. 3, pp. 496–501, 2005.

[11] J. Tepperman and S. S. Narayanan, "Better nonnative intonation scores through prosodic theory," in *INTERSPEECH 2008*, 2008, pp. 1813–1816.

[12] T. Cincarek, R. Gruhn, C. Hacker, E. Noth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-natives first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65 – 88, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230808000193

[13] J. Cheng, "Automatic assessment of prosody in high-stakes English tests." in *INTERSPEECH 2011*. ISCA, 2011, pp. 1589–1592.

[14] F. Hönig, A. Batliner, K. Weilhammer, and E. Noth, "Automatic assessment of non-native prosody for English as L2," in *Speech Prosody*, 2010.

[15] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, pp. 265–292, 1999.

[16] V. Dellwo and P. Wagner, "Relations between language rhythm and speech rate," in *ICPhS*, 2003, pp. 471–474.

[17] E. Grabe and E. Low, "Durational Variability in Speech and the Rhythm Class Hypothesis," in *Laboratory Phonology VII*, 2002, pp. 515–546.

[18] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[19] T. Giorgino *et al.*, "Computing and visualizing dynamic time warping alignments in r: the dtw package," *Journal of statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[20] A. Sakurai, K. Hirose, and N. Minematsu, "Data-driven generation of f 0 contours using a superpositional model," *Speech Communication*, vol. 40, no. 4, pp. 535–549, 2003.

[21] D. Escudero-Mancebo, C. Gonzlez-Ferreras, L. Aguilar, and E. Estebas-Vilaplana, "Automatic assessment of non-native prosody by measuring distances on prosodic label sequences," in *Proc. Interspeech 2017*, 2017, pp. 1442–1446.