Automatic assessment of non-native prosody by measuring distances on prosodic label sequences

David Escudero-Mancebo¹, César González-Ferreras¹, Lourdes Aguilar², Eva Estebas-Vilaplana³

¹Department of Computer Science, University of Valladolid, Spain
²Department of Spanish Philology, Universitat Autònoma de Barcelona, Spain
³Department of Modern Languages, UNED, Spain
descuder@infor.uva.es

Abstract

The aim of this paper is to investigate how automatic prosodic labeling systems contribute to the evaluation of non-native pronunciation. In particular, it examines the efficiency of a group of metrics to evaluate the prosodic competence of non-native speakers, based on the information provided by sequences of labels in the analysis of both native and non-native speech. A group of SpToBI labels were obtained by means of an automatic labeling system for the speech of native and non-native speakers who read the same texts. The metrics assessed the differences in the prosodic labels for both speech samples. The results showed the efficiency of the metrics to set apart both groups of speakers. Furthermore, they exhibited how non-native speakers (American and Japanese speakers) improved their Spanish productions after doing a set of listening and repeating activities. Finally, this study also shows that the results provided by the metrics are correlated with the scores given by human evaluators on the productions of the different speakers.

Index Terms: Prosody and second language, computer assisted pronunciation training, Prosodic ToBI labeling

1. Introduction

Computer assisted pronunciation training (CAPT) systems have shown to be attractive both from a pedagogical and a commercial point of view. These systems mainly focus on the training of phonetic pronunciation, paying less attention to prosodic aspects (with the only exception of fluency). Nevertheless, prosody plays an important role in the evaluation protocols of L2 evaluators; for example, [1] establishes the minimum requirements of prosodic competence to assess the level of Spanish proficiency according to the Common European Framework of Reference for Languages (CEFR).

There are several approaches in the state of the art that face up the problem of evaluating L2 prosody [2]. These systems are based on comparing the prosodic acoustic characteristics of L2 utterances (like F0, duration and energy) with the corresponding features of native speakers (generally with the ones of a golden speaker who is considered to use the correct pronunciation). These approaches have an important limitation that has to do with the under representation of variety in prosody: the same prosodic function can be represented with more than one prosodic form [3]. This is challenging for CAPT systems because two prosodic productions of the same text can be different but valid at the same time. To face up this problem, in this work we have devised a double strategy: on the one hand, we have used prosodic labels (no directly prosodic acoustic features) to compare utterances; on the other hand, L2 utterances have not only been compared with those of a single golden speaker but with the productions of a set of reference speakers.

The efficiency of using prosodic labels (a set of symbols for transcription the intonation patterns and other aspects of the prosody of utterances) has been well established in the context of L2 assessment [4, 5, 6]. Related to this, the ToBI system is a broadly accepted framework for the transcription of prosodic phenomena. It was originally developed for English, based on Pierrehumberts autosegmental model, but since then it has been applied to a large number of languages, among them Spanish [7]. In [4], an experiment of style identification was presented by using the Automatic ToBI labels described in [8]: the results showed 95% of accuracy. When a given utterance is labeled with prosodic labels, its representation is simplified since the labels include symbolic information that specifies the relevant prosodic functions present in the utterance. The automatic prosodic labeling systems are prepared to process prosodic variety as they are trained with data that reflects the form-function multiplicity. In [9], we used the automatic SpToBI classifier presented in [10] to characterize radio broadcasting prosodic style by measuring the mutual information between sequences of prosodic labels. In this paper, we follow a similar approach to compute distances between native and non-native speakers by improving the mutual information metric used in [9] by applying normalization that takes into account the joint entropy of the labels of the different type of speakers. The results show that these new metrics permit to identify non-native speakers with a degree of confidence that is statistically significant. The results are consistent with the a-priori expected improvement on the pronunciation as the pronunciation exercises are repeated.

Unlike other studies that compare L2 prosodic contours with those of a single golden speaker, in this work we use a group of native speakers (as a whole, not individually) as reference (already done in [11]). This is well motivated by the previous research in [9], where we showed the high variety that could exist between speakers of the same style when reading the same texts. This fact evidences the limitations of comparing F0 utterances with the ones of a single native speaker. The aim of this new study is to demonstrate that using the minimum and/or maximum distance between the L2 utterances of non-native speakers and the corresponding native utterances permits to obtain a better correlation between the objective quality metrics and the subjective scores assigned by human evaluators.

Section 2 details the experimental procedure presenting the
corpus, the automatic prosodic labeling systems and the metrics. Section 3 presents the results and the paper ends with a discussion about the potential of prosodic labels to offer information on the limitations of non-native speaker pronunciation. Section 4 discusses the prosodic variability found in Spanish [12].

The whole SAMPLE corpus is described in [13]. It contains different materials: sentences, the Aesops Fable “The North Wind” and news paragraphs. In this study, we focus on the sentences. They were extracted from the news paragraphs of the Glissando corpus [12]. The list of sentences is described in [13] (see table 1 of that paper). All sentences followed a phonetic coverage criterion. 14 speakers that were students of Spanish were recorded: 9 American English (AM) and 5 Japanese (JP). All of them were students of Spanish at a university level. In this paper, we refer to the American speakers as AM1, AM2, AM3, AM4, AM5, AM6, AM7, AM8 and AM9, (corresponding to f01, f02, f04, f05, f06, f07, m08, f09 and f10 in the SAMPLE corpus), where f means female and m means male. Similarly, Japanese speakers are referred to as JP1, JP2, JP3, JP4 and JP5 (corresponding to m03, f11, f12, f13 and f14 in the SAMPLE corpus).

There were several repetitions of each of the fifteen sentences (s01-s15). Ten sentences (s01-s10) were read three times by the L2 speakers. Another group of ten sentences (s06-s15) were used for the task of listen and repeat: a reference utterance immediately afterwards. This task was recorded three times. Therefore, we can define six blocks of sentences:

- BR1: read sentences s01-s10.
- BR2: read sentences s01-s10.
- BR3: read sentences s01-s10.
- BLR1: listen and repeat sentences s06-s15.
- BLR2: listen and repeat sentences s06-s15.
- BLR3: listen and repeat sentences s06-s15.

The reference sentences of native pronunciation are the corresponding fifteen sentences extracted from the Glissando corpus. As the Glissando corpus recorded eight different professional speakers, we have more than one reference to contrast the non-native pronunciation. In this paper, the professional speakers are referred to as SP1, SP2, SP3, SP4, SP5, SP6, SP7 and SP8 (corresponding to f10a, f11r, f13r, f15a, m09a, m10a, m12r and m14r in the Glissando corpus). As before, f means female and m means male. Furthermore, r stands for a radio speaker and a indicates an actor. In [13], the procedure followed for subjective evaluation of the utterances of the corpus is described. As a result of this procedure, all the speakers obtained different numeric scores representing the quality of his/her pronunciation taking into account different aspects that have to do with both phonetic and prosodic pronunciation proficiency.

2.2. Automatic prosodic labeling

For the labeling of the spoken material, the procedure described in [14] was used. An automatic labeling system was trained with a subcorpus of the Glissando corpus consisting of a 60 news items recorded by five professional speakers (12 news each). These news items include a total of 5,103 pitch accents and 2,835 boundary tones.

The automatic system is a pairwise coupling classifier that combines evidences of three complementary types of classifiers, such as artificial neural networks (NN), decision trees (DT), and support vector machines (SVM) [10]. In order to combine the three classification modules (DT, NN and SVM), we used the comprehensive fuzzy technique proposed in [15].

The reference unit for the automatic labeling system is the word. Every word is characterized in terms of prosodic information (F0, energy and duration features) and POS tags, as described in [10]. As a result, we obtain up to two SpToBI labels per word: one for the pitch accent and another one for the boundary tone. We use the following SpToBI pitch accents: H+, L* = {L+ U L*+H U H+L*}, L+>H*, L+H* = {L+H* U (L+)>H*}, L+H* = {L+H* U (L+)>H*}. H+, L* = {L* U L*+H U H+L*}, L+>H*, L+H* = {L+H* U (L+)>H*}, and the following boundary tones: L%, H%, =%, 'H%', LH% = [LH% U LH%]. Additionally, the label "none" represents the absence of tone.

2.3. The metrics

The output of the automatic labeling is a sequence of prosodic labels per utterance. By comparing the sequences of the automatic labels that correspond to different speakers, we should obtain a clue of the similarity of the prosodic productions of both speakers. By computing the mutual information between the sequences of prosodic labels of two speakers (as in [9]), we obtain a value that indicates the quantity of information that the speakers share. As the speakers read the same text, the prosodic sequences of the different speakers should have similar informational content. We use in this paper metrics based on the mutual information between sequences of labels of native and non-native speakers as a measure of the pronunciation quality.

Mutual information is defined as:

\[ I(X; Y) = H(X) + H(Y) - H(X, Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (1) \]

Being x and y the prosodic labels of the utterances read by the speaker X and Y respectively. The higher the similarity between the sequences of labels, the higher the value of \( I(X; Y) \).

A variant of mutual information named variation of information [16] satisfies the properties of a metric (triangle inequality, non-negative, indiscernibility and symmetry):

\[ d(X, Y) = H(X, Y) - I(X; Y) \quad (2) \]
which can be normalized as:

\[ D(X, Y) = \frac{d(X, Y)}{H(X, Y)} = 1 - I(X; Y)/H(X, Y) \]  

(3)

A normalized version of the mutual information is

\[ I'(X; Y) = I(X; Y)/\min[H(X), H(Y)] \]  

(4)

In the case of \( d(X, Y), d(Y, X) \) and \( I'(X; Y) \), the closer the value to zero, the more similar \( x \) and \( y \) are.

3. Results

Table 1 shows the \( d(X, Y) \) distances between the different speakers of the corpus (native and non-native) with respect to the native ones. The general tendency is that distances between non-native speakers and native speakers are higher than distances between native speakers. Thus, for example, in column SP8, the distances corresponding to non-native speakers ranges from 1.94 to 2.68 whereas the distances corresponding to native speakers are between 1.871 and 1.939. This tendency is magnified when the \( \min \) and \( \max \) columns are analyzed: The mean value of the values of column \( \min \) for non-native speakers (AM1..JP5) measures 1.93 and 1.71 for native speakers (SP1..SP8 rows). In the \( \max \) column, the mean value for non-native speakers is 2.20 and it is 1.99 for native speakers.

Table 2 shows the mean values and confidence intervals of the distances between groups of users computed by using the different metrics described in section 2.3 applied to the different blocks of sentences detailed in section 2.1. \( \mu \) is the mean value and \( CI \) is the confidence interval.

### Table 1: Variation of information \( d(X, Y) \) between the different speakers with respect to the native speakers (columns SP1..SP8). Columns \( \min \) and \( \max \) are the minimum and maximum value of columns SP1..SP8 in each row. In the rows corresponding to non-native speakers zero values (the distance of the speaker to herself) are omitted to compute \( \min \) and \( \max \) values. The sentences of the block BR1 have been used for the computation of this table.

| Block | SP1 | SP2 | SP3 | SP4 | SP5 | SP6 | SP7 | SP8 | \( \min \) | \( \max \) |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|---------|---------|
| BR1   | 2.02 | 2.03 | 2.03 | 2.02 | 2.06 | 2.04 | 2.34 | 2.33 | 2.04    | 2.33    |
| BR2   | 2.08 | 2.09 | 2.08 | 2.07 | 2.05 | 2.07 | 2.32 | 2.31 | 2.09    | 2.31    |
| BR3   | 2.13 | 2.13 | 2.13 | 2.13 | 2.13 | 2.13 | 2.32 | 2.32 | 2.13    | 2.32    |
| BR4   | 2.18 | 2.17 | 2.17 | 2.17 | 2.17 | 2.17 | 2.32 | 2.32 | 2.17    | 2.32    |
| BR5   | 2.23 | 2.23 | 2.23 | 2.23 | 2.23 | 2.23 | 2.32 | 2.32 | 2.23    | 2.32    |
| BR6   | 2.28 | 2.28 | 2.28 | 2.28 | 2.28 | 2.28 | 2.32 | 2.32 | 2.28    | 2.32    |
| BR7   | 2.33 | 2.33 | 2.33 | 2.33 | 2.33 | 2.33 | 2.32 | 2.32 | 2.33    | 2.32    |
| BR8   | 2.38 | 2.38 | 2.38 | 2.38 | 2.38 | 2.38 | 2.32 | 2.32 | 2.38    | 2.32    |

3. Results

Table 1 shows the \( d(X, Y) \) distances between the different speakers of the corpus (native and non-native) with respect to the native ones. The general tendency is that distances between non-native speakers and native speakers are higher than distances between native speakers. Thus, for example, in column SP8, the distances corresponding to non-native speakers ranges from 1.94 to 2.68 whereas the distances corresponding to native speakers are between 1.871 and 1.939. This tendency is magnified when the \( \min \) and \( \max \) columns are analyzed: The mean value of the values of column \( \min \) for non-native speakers (AM1..JP5) measures 1.93 and 1.71 for native speakers (SP1..SP8 rows). In the \( \max \) column, the mean value for non-native speakers is 2.20 and it is 1.99 for native speakers.

Table 2 shows the mean values and confidence intervals of the distances between groups of users computed by using the different metrics described in section 2.3 applied to the different blocks of sentences detailed in section 2.1. \( \mu \) is the mean value and \( CI \) is the confidence interval.

| Block | No native speakers | Native speakers |
|-------|--------------------|-----------------|
| \( \mu \) | \( 95\% \ CI \) | \( \mu \) | \( 95\% \ CI \) |
| BR1   | 1.430 [1.417, 1.444] | 1.631 [1.551, 1.712] |
| BR2   | 1.435 [1.419, 1.451] | 1.631 [1.551, 1.712] |
| BR3   | 1.443 [1.427, 1.458] | 1.631 [1.551, 1.712] |
| BLR1  | 1.413 [1.399, 1.427] | 1.615 [1.534, 1.695] |
| BLR2  | 1.426 [1.412, 1.440] | 1.615 [1.534, 1.695] |
| BLR3  | 1.440 [1.427, 1.454] | 1.615 [1.534, 1.695] |

Additionally, distances in table 2 show a tendency to decrease (increase in the case of \( I'(X, Y) \)) when the reading activities are repeated; for example the metric \( D(X, Y) \) is 0.592 for block BR1 and 0.588 for block BR3. Again, values are generally smaller in the reading after the listening activities: for example, the metric \( I'(X, Y) \) is 0.432 for BR1 and 0.429 for BLR1.

The mean values in table 2 exhibit that normalized versions of the metrics \( D(X, Y) \) and \( I'(X, Y) \) show the highest degree of consistency so that \( \mu(BR1) > \mu(BR2) > \mu(BR3); \mu(BLR1) > \mu(BLR2) > \mu(BLR3) \) and \( \mu(BRi) > \mu(BLRi) \) for \( i = 1, 2, 3 \).

Table 3 presents the correlation between the subjective scores assigned to the speakers by human evaluators and the objective distance between the evaluated speakers and the reference native ones. We select the scores assigned to prosodic-related variables (Fluency, Accent, Rhythm) and the overall evaluation score named DELE. The correlation ranges from 0.39 to 0.53 in all the cases. This correlation increases when the \( \min \) and \( \max \) rows are analyzed. In this case, \( \min \) and \( \max \) indicate, respectively, the correlation between the subjective score and the minimal or maximum distance to any reference native speaker. The intervals range from 0.62 to 0.66 for \( \mu \) but it is 0.79 for \( max \).
Table 4: Automatic prosodic labels obtained from the different speakers’ utterances of the sentence “La coalicién interpuso esta querella por prevaricacin el viernes pasado” (The coalition interposes this complaint for prevarication last friday).

| Word               | SP1   | SP2   | SP3   | SP4   | SP5   | SP6   | SP7   | SP8   | JP1   |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| la                 | H*    | L+H*  | H*    | H*    | L+H*  | L+H*  | H*    | H*    | H*    |
| coalicién         | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  |
| interpuso          | H*    | L+H*  | H*    | H*    | L+H*  | L+H*  | H*    | H*    | H*    |
| esta               | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  |
| querella           | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  |
| por                | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  |
| prevaricacin       | H*    | L+H*  | H*    | H*    | L+H*  | L+H*  | H*    | H*    | H*    |
| el                 | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  |
| viernes            | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  |
| pasado             | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  | L+H*  |

Table 3: Correlation between objective and subjective scores for the speakers in the corpus.

| Reference | Fluency | Accent | Rythm | DELE |
|-----------|---------|--------|-------|------|
| SP1       | -0.39   | -0.42  | -0.44 | -0.41|
| SP2       | -0.46   | -0.45  | -0.48 | -0.47|
| SP3       | -0.40   | -0.43  | -0.41 | -0.42|
| SP4       | -0.42   | -0.46  | -0.43 | -0.44|
| SP5       | -0.44   | -0.48  | -0.47 | -0.46|
| SP6       | -0.50   | -0.49  | -0.53 | -0.50|
| SP7       | -0.45   | -0.46  | -0.49 | -0.45|
| SP8       | -0.44   | -0.48  | -0.41 | -0.46|
| min       | -0.62   | -0.64  | -0.66 | -0.63|
| max       | -0.79   | -0.79  | -0.79 | -0.79|

4. Discussion

The results show that the use of mutual information as a distance measure between speakers, as found in [9], is not the best option in this scenario. On the contrary, it is necessary to consider joint entropy and/or normalize results to increase the reliability of the results.

The four metrics that have been proved in this study are useful to show the separation between the two groups of speakers (native and non-native), and the normalized metrics properly cover the improvements after repetitions.

The results highlight the risks of using a single speaker as a reference speaker when assessing the quality of non-native speaker prosody. Such result was expected, since it is well known that a same sentence can be pronounced with different prosody by different speakers being all these valid pronunciations. To take into account the prosodic variety and the diversity of possible locations, it may be advantageous to take into account the whole set of reference speakers instead of a single golden speaker. In this paper, we have shown that using the whole set of reference speakers instead of a single one, the use of possible locutions, it may be advantageous to take into account the prosodic variety and the diversity of possible locations.

5. Conclusions and future work

In this work we have presented the use of a set of metrics based on joint entropy for computing distances between sequences of prosodic labels. These metrics have shown to be efficient to discriminate native from non-native utterances. It has also been shown that the metrics correlate with the subjective scores of quality and that the computed distances are consistent with respect to the expected results after the repetition exercises.

In future work, we will examine the combination of the metrics with other possible complementary metrics that permit to increase the results for automatic assessment of the pronunciation quality. We will also work on the development of a module for the diagnosis of the pronunciation deficits that benefits from the expressiveness of the ToBI labels as a standard for representing the relationship between prosodic form and function.

6. References

[1] A. G. Santa-Cecilia, “Plan curricular del instituto cervantes: niveles de referencia para el espaol,” MarcoELE: Revista de didtica, no. 5, p. 1, 2007.
[2] J. P. Arias, N. B. Yoma, and H. Vivanco, “Automatic intonation assessment for computer aided language learning,” Speech communication, vol. 52, no. 3, pp. 254–267, 2010.
[3] D. Hirst, “Form and function in the representation of speech prosody,” Speech Communication, vol. 46, no. 34, pp. 334 – 347, 2005.
[4] A. Rosenberg, “Symbolic and direct sequential modeling of prosody for classification of speaking-style and nativeness.” in INTERSPEECH, 2011, pp. 1065–1068.

[5] J.-m. Kim, “Annotation of a non-native english speech database by korean speakers.” Speech Sciences, vol. 9, no. 1, pp. 111–135, 2002.

[6] J. Tepperman, A. Kazemzadeh, and S. Narayanan, “A text-free approach to assessing nonnative intonation.” in INTERSPEECH, 2007, pp. 2169–2172.

[7] P. Prieto, Transcription of intonation of the Spanish language. Lincom Europa, 2010.

[8] A. Rosenberg, “Autobi-a tool for automatic tobi annotation.” in Interspeech, 2010, pp. 146–149.

[9] D. Escudero, C. González, Y. Gutiérrez, and E. Rodero, “Identifying characteristic prosodic patterns through the analysis of the information of sp_tobi label sequences,” Computer Speech and Language, vol. 45, pp. 39 – 57, 2017.

[10] C. Gonzalez-Ferreras, D. Escudero-Mancebo, C. Vivaracho-Pascual, and V. Cardeñoso Payo, “Improving automatic classification of prosodic events by pairwise coupling,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 7, pp. 2045–2058, sept. 2012.

[11] H. Ding, R. Hoffmann, and D. Hirst, “Prosodic transfer: A comparison study of f0 patterns in l2 english by chinese speakers,” in Speech Prosody, 2016, pp. 756–760.

[12] J.-M. Garrido, D. Escudero, L. Aguilar, V. Cardeñoso, E. Rodero, C. de-la Mota, C. González, C. Vivaracho, S. Rustullet, O. Larrea, Y. Laplaza, F. Vizcaíno, E. Estebas, M. Cabrera, and A. Bonafonte, “Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan,” Language Resources and Evaluation, vol. 47, no. 4, pp. 945–971, 2013.

[13] 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). Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014, pp. 3967–3972.

[14] D. Escudero, L. Aguilar, C. González, V. Cardeñoso, and Y. Gutiérrez, “Preliminary results on SpToBI prosodic labeling assisted by an automatic fuzzy classifier,” in Proceedings of the 7th International Conference on Speech Prosody, Dublin, Ireland, May 2014, pp. 457–461.

[15] D. Escudero-Mancebo, C. González-Ferreras, C. Vivaracho-Pascual, and V. Cardeñoso Payo, “A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling,” Computer Speech and Language, vol. 28, no. 1, pp. 326 – 341, 2014.

[16] M. Meilà, “Comparing clusterings by the variation of information,” in Learning theory and kernel machines. Springer, 2003, pp. 173–187.