INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

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

David Escudero-Mancebo[1], *César González-Ferreras*[1],
Lourdes Aguilar[2], *Eva Estebas-Vilaplana*[3]

[1]Department of Computer Science, University of Valladolid, Spain
[2]Department of Spanish Philology, Universitat Autònoma de Barcelona, Spain
[3]Department 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 Sp_ToBI 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 pronuntiation trainning, Prosodic ToBI labeling [1]

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 transcribing 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 Sp_ToBI 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] and 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 singlne 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

[1]Thanks to: Research project TIN2014-59852-R (MIMECO/FEDER, UE) VIDEOJUEGOS SOCIALES PARA LA ASISTENCIA Y MEJORA DE LA PRONUNCIACION DE LA LENGUA ESPAÑOLA.

http://dx.doi.org/10.21437/Interspeech.2017-366

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 corpus, the automatic labeling systems and the metrics.

2. Experimental procedure

2.1. The Corpus

In the framework of the SAMPLE research project, a corpus of spoken utterance produced by L2 Spanish non-native speakers was developed as a means to support computer-assisted pronunciation training (CAPT) studies. The central part of the corpus includes a set of sentences and paragraphs selected from the news database of a popular Spanish radio news broadcasting station. The texts cover various information domains related to everyday life. They were obtained from the Glissando corpus [12], which was developed in connection to another project related to automatic prosodic labelling. The materials belong to the subset of prosodically balanced sentences in Glissando, which statistically resemble 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 of each sentence by a native professional speaker was presented to the non-native speakers, who had to listen and repeat it 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 f16a, 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 Sp_ToBI labels per word: one for the pitch accent and another one for the boundary tone. We use the following Sp_ToBI pitch accents: H*, L* = {L* ∪ L*+H ∪ H+L* }, L+>H*, L+H* ={L+H* ∪ (L+)H*}, L+!H* ={L+!H* ∪ (L+)!H* ∪ !H*}, L+H* ={L+H* ∪ (L+)H* ∪ H*}; and the following boundary tones: L%, H%, =%, !H%, LH% ={LH% ∪ L!H%}. 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 two 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 test, 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:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \quad (1)$$

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, indiscernability and symmetry):

$$d(X,Y) = H(X,Y) - I(X;Y) \quad (2)$$

Table 1: *Variation of informacin $d(X,Y)$ 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.*

	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8	min	max
AM1	1.92	2.21	2.15	2.03	2.02	2.06	2.04	2.11	1.92	2.21
AM2	2.09	2.14	2.21	2.10	2.17	2.09	2.13	2.15	2.09	2.21
AM3	2.18	2.26	2.19	2.15	2.12	2.17	1.93	2.28	1.93	2.28
AM4	1.98	2.06	2.05	2.11	2.11	2.09	2.09	2.07	1.98	2.11
AM5	1.69	2.11	2.00	1.92	1.90	1.84	1.96	2.11	1.69	2.11
AM6	2.04	2.10	2.20	1.99	1.92	2.11	2.03	1.97	1.92	2.20
AM7	2.13	2.24	2.25	2.29	2.01	2.18	2.18	2.12	2.01	2.29
AM8	2.00	2.21	2.08	1.91	2.02	1.88	2.02	2.09	1.88	2.21
AM9	1.77	1.92	1.99	1.91	1.77	1.83	1.91	1.94	1.77	1.99
JP1	1.87	2.11	2.05	1.97	1.96	1.88	1.97	2.01	1.87	2.11
JP2	2.27	2.29	2.22	2.26	2.31	2.12	2.12	2.23	2.12	2.31
JP3	1.89	2.29	2.25	2.09	2.22	2.20	2.02	2.20	1.89	2.29
JP4	1.96	2.22	2.21	2.02	1.98	1.99	1.94	2.17	1.94	2.22
JP5	2.07	2.22	2.21	2.16	2.05	2.24	2.21	2.09	2.05	2.24
SP1	0.00	2.02	1.86	1.62	1.77	1.61	1.71	1.88	1.61	2.02
SP2	2.02	0.00	2.06	1.93	1.93	1.89	1.86	1.89	1.86	2.06
SP3	1.86	2.06	0.00	1.79	2.02	1.88	2.01	1.90	1.79	2.06
SP4	1.62	1.93	1.79	0.00	1.66	1.87	1.92	1.92	1.62	1.93
SP5	1.77	1.93	2.02	1.66	0.00	1.81	1.74	1.89	1.66	2.02
SP6	1.61	1.89	1.88	1.87	1.81	0.00	1.72	1.80	1.61	1.89
SP7	1.71	1.84	2.00	1.91	1.74	1.71	0.00	1.94	1.71	2.00
SP8	1.88	1.89	1.90	1.92	1.89	1.80	1.94	0.00	1.80	1.94

which can be normalized as:

$$D(X,Y) = d(X,Y)/H(X,Y) = 1 - I(X;Y)/H(X,Y) \quad (3)$$

A normalized version of the mutual information is

$$I'(X;Y) = I(X;Y)/\min[H(X), H(Y)] \quad (4)$$

In the case of $d(X,Y)$, $d(X,Y)$ 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 cells of the distance tables like 1 computed by using the four metrics detailed in section 2.3, applied to the six blocks of sentences detailed in section 2.1 (the whole 24 tables are not presented for the lack of space). The table compares the statistics between native and non-native speakers. As the native speakers did not do the repetitions, the values corresponding to BR1, BR2 and BR3 and the values corresponding to BLR1, BLR2 and BLR3 are the same. The four metrics show significant statistical differences between non-native and native speakers in all the blocks when the t-student test is applied

Table 2: *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. μ is the mean value and CI is the confidence interval.*

	Block	No native speakers		Native speakers	
		μ	95% CI	μ	95% CI
I(X;Y)	BR1	1.430	[1.417, 1.444]	1.631	[1.551, 1.711]
	BR2	1.435	[1.419, 1.451]	1.631	[1.551, 1.711]
	BR3	1.443	[1.427, 1.458]	1.631	[1.551, 1.711]
	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]
d(X,Y)	BR1	2.076	[2.052, 2.100]	1.852	[1.821, 1.883]
	BR2	2.066	[2.044, 2.088]	1.852	[1.821, 1.883]
	BR3	2.057	[2.037, 2.078]	1.852	[1.821, 1.883]
	BLR1	2.041	[2.014, 2.068]	1.874	[1.839, 1.908]
	BLR2	1.986	[1.960, 2.011]	1.874	[1.839, 1.908]
	BLR3	1.989	[1.962, 2.015]	1.874	[1.839, 1.908]
D(X,Y)	BR1	0.592	[0.587, 0.596]	0.550	[0.543, 0.557]
	BR2	0.590	[0.585, 0.595]	0.550	[0.543, 0.557]
	BR3	0.588	[0.583, 0.592]	0.550	[0.543, 0.557]
	BLR1	0.590	[0.585, 0.596]	0.555	[0.548, 0.563]
	BLR2	0.582	[0.576, 0.587]	0.555	[0.548, 0.563]
	BLR3	0.579	[0.574, 0.585]	0.555	[0.548, 0.563]
I'(X,Y)	BR1	0.432	[0.427, 0.437]	0.392	[0.386, 0.399]
	BR2	0.430	[0.425, 0.435]	0.392	[0.386, 0.399]
	BR3	0.429	[0.424, 0.434]	0.392	[0.386, 0.399]
	BLR1	0.429	[0.424, 0.435]	0.396	[0.389, 0.404]
	BLR2	0.421	[0.416, 0.426]	0.396	[0.389, 0.404]
	BLR3	0.416	[0.411, 0.422]	0.396	[0.389, 0.404]

with $p - value << 0.001$. Smaller values for native speakers (higher for $I(X,Y)$) indicate that the similarity between the native speakers is higher than the similarity between non-native and native speakers.

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 min but it is 0.79 for max.

Table 4: *Automatic prosodic labels obtained from the different speakers' utterances of the sentence "La coalicin 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*	L+>H*		
coalición	L+H* H%	L+H* !H%	L+H* H%	L+H* =%	L+H* =%	L+>H*	L+H* =%	L+H* H%	L+H* !H%
interpuso	L+H* !H%	L+H*	L+H* !H%	L+H* H%	L+H*	L+H*	L+H*	L+>H*	L+H* LH%
esta	L+>H*		L+>H*	L+H*	L+H*				L+H* LH%
querella	L+H* LH%	L+!H* H%	L+H* H%	L+H* LH%	L+H* H%	L+H* =%	L+H* H%	L+!H* H%	L+H*
por			H*	L+H*	H*				H*
prevaricación	L+H* LH%	L+!H* H%	L+!H* !H%	L+H* H%	L+H* !H%	L+H* !H%	L+H* H%	L+H* H%	L+H* LH%
el	H*					H*			H*
viernes	L+H*		L+>H*	L+H*	L+H*	L+!H*	L+!H*	L+!H*	L+H*
pasado	L* L%	L+H*	L+H* !H%	L+H* L%	L*	L+H* L%	L+H* L%	L+>H*	L* L%

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 intonations by different speakers being all these valid pronunciations.

To take into account the prosodic variety and the diversity of possible locutions, 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 closest or the most distant speaker as the selection criterium is effective. However, other agglutination scores will be tested in future work.

The example in table 5 illustrates why the measures based on mutual information work. The sequences of native speakers (SP1 to SP8) have more similarities between them (and thus more mutual information) than with the non-native speaker's sequence (JP1). The most revealing differences concern to the presence/absence of pitch accent and the location of boundary tones. The monosyllabic functional words "el" (the) and "por" (for) have been accented by the non-native speaker with a high tone H*, whereas none native speaker has placed an accent on these words. With respect to boundary tones, at the beginning of the sentence, the non-native speaker clearly shows a preference for short prosodic groups "la coalición / interpuso" (the coalition / interposed) and makes a prosodic mistake with the insertion of a boundary tone after the functional word "esta" (this). This violates the good formation of prosodic groups in Spanish, since "esta" operates as a clitic word. Contrary to this phrasing, the native speakers coincide to segment the sentence after "querella" (complaint).

As far as pitch accents are concerned, the inventory has been reduced in the non-native pronunciations, since neither the default value in Spanish prosody L+>H*, a rising accent with a peak displacement, nor L+!H*, a downstepped rising accent without peak displacement, appear in the data. On the contrary, the final rising accent (LH%) which is less used by native speakers, frequently appear in the non-native pronunciation. This can be a case of prosodic transference. We are currently working on the use of these evidences to identify prosodic mistakes in order to obtain a diagnosis of the specific problems of each speaker that allows us to give indications for further improvement.

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 pronuntiation 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 español," *MarcoELE: Revista de didáctica*, 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*, 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 Sp_ToBI 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. Meilă, "Comparing clusterings by the variation of information," in *Learning theory and kernel machines*. Springer, 2003, pp. 173–187.