Cross-lingual English Spanish tonal accent labeling using decision trees and neural networks

David Escudero-Mancebo[1], Lourdes Aguilar[2], César González Ferreras[1], Carlos Vivaracho Pascual[1], and Valentín Cardeñoso-Payo[1]

[1] Dpt. of Computer Sciences, Universidad de Valladolid, Spain
[2] Dpt. of Spanish Philology, Universidad Autónoma de Barcelona, Spain

Abstract. In this paper we present an experimental study on how corpus-based automatic prosodic information labeling can be transferred from a source language to a different target language. The Spanish ESMA corpus is used to train models for the identification of the prominent words. Then, the models are used to identify the accented words of the English Boston University Radio News Corpus (BURNC). The inverse process (training the models with English data and testing with the Spanish corpus) is also contrasted with the results obtained in the conventional scenario: training and testing using the same corpus. We got up to 82.7% correct annotation rates in cross-lingual experiments, which contrast slightly with the accuracy obtained in a mono-lingual single speaker scenarios (86.6% for Spanish and 80.5% for English). Speaker independent monolingual recognition experiments have been also performed with the BURNC corpus, leading to cross-speakers results that go from 69.3% to 84.2% recognition rates. As these results are comparable to the ones obtained in the cross-lingual scenario we conclude that the new approach we defend has to face up with similar challenges as the ones presented in speaker independent scenarios.

Index Terms: prosodic labeling, cross-lingual prosody, automatic accent identification [3]

1 Introduction

The prosodic function of emphasis or accent is used to make some parts of the spoken message like words, syllables, turn... more relevant with respect to the rest of the message. The identification or labeling of the accented words in a given message has several practical applications in diverse fields of spoken technologies. Thus, in *speech recognition* the accents can be useful to disambiguate confusing words like *sé* verb vs. *se* relative in Spanish; in *text to speech* because the modeling of the accents improves the naturalness of the synthetic output;

[3] This work has been partially supported by Consejería de Educación de la Junta de Castilla y León project VA322A11-2 and by Ministerio de Ciencia e Innovacion, Spanish Government Glissando project FFI2008-04982-C003-02

Corpus	ESMA-UPC	BURNC	BURNC.f1a	BURNC.f2b	BURNC.f3a	BURNC.m1b	BURNC.m2b	BURNC.m3b
#words	7236	27767	3790	11994	2624	3974	3413	1972
#accented	2483	13899	2053	6214	1281	1604	1823	924
#un-accented	4895	14586	1831	6057	1438	2467	1700	1093

Table 1. Number of words in the corpora and subcorpora

in *speaker recognition*, because the tonal accents and boundary tones represent the most characteristic pitch movements of a given speaker or group of speakers; and in *dialog systems* because the different turns can be characterized according to the number and type of accents that the speaker or the machine use.

There are several approximations to the automatic identification of accents in the state of the art. [1] or [12] are two good examples, not only because they reach identification rates that are close to 90%, but also because they make an extensive review of the state of the art. Acoustic, lexical and syntactic evidences are combined in a maximum entropy framework to predict the tonal accents, the boundary tones and the breaks in the Boston University News Radio Corpus[10]. Despite the prediction must be reviewed by manual transcribers of prosody, the benefits are important because the manual ToBI labeling is estimated to take from 100-200 times the real time [13].

Here we explore a cross-lingual approach where a given corpus with prosodic labels will be used to predict the labels of a different corpus in a different language. Despite the shape of the accents is highly dependent on the language, the emphatic function is universal and it is dependent of the same acoustic cues: the variation of the acoustic prosodic features in a given unit with respect to the context.

Under this hypothesis, the Boston University Radio News Corpus is used to train non linear models (decision trees and neural networks) that are then used to identify the presence of tonal accent in a Spanish corpus and vice versa. Results are promising as more than 80% of the words are correctly classified. The paper shows that these rates are similar to the ones we obtain when the different speakers of the Boston corpus are contrasted with each other following the same approach: the data of a single speaker is used to identify the tonal accent presence of the others. First, the experimental procedure is presented, followed by the results on cross-lingual accent identification. Discussion and future work end the paper.

2 Experimental procedure

The cross-lingual approach consists of training with the data of a corpus in a given language and testing with data of a different language. Cross-lingual differences are contrasted with cross-speaker differences as the BURNC permits training and testing using the data of the six speakers separately. Additionally, we systematically contrast the differences on the input features among the diverse corpora in several practical aspects.

First, the scale of the input features is analyzed to contrast the differences among the languages and speakers. The cross-speakers and/or cross-lingual accent identification task shows the clear impact of the scale variability on the performance of the classifiers, which justifies the need of the normalization process.

The cross lingual study continues with the examination of the relevance of the input features in the different corpora. Input features are ranked in terms of their informative capabilities for discriminating whether a word or a syllable is accented. Every language and speaker has its own ranking to be contrasted. Furthermore, the most informative input features are also analyzed to verify whether significant differences appear among the diverse corpora (e.g. f0_range of the accents in Spanish vs. accents in English).

Finally, automatic prediction results are contrasted with perceptual judgements made by a set of labelers on the same testing corpus. This is useful to value the usefulness of the automatic labeling process facing its application in prosodic labeling of corpora.

2.1 Processing of the corpora

The Boston University Radio News Corpus BURNC [10] has been used to model English tonal accents in this work. This corpus includes labels separating phonemes, syllables and words. Accents are marked with a ToBI label and a position. Inspired in state of the art works [15, 1], the accent tones were aligned with respect to the prominent syllable and the word containing it (words with more than one label are discarded in this work). All the utterances in the corpus with ToBI labels, from all the speakers (females: f1a, f2b, f3a; males: m1b, m2b and m3b) have been used, as shown in Table 1.

The Spanish corpus used in this paper was ESMA-UPC. It was designed for the construction of a unit concatenative TTS system for Catalan and Spanish at the UPC (http://www.gps-tsc.upc.es) [2]. It contains recordings of three hours of spoken utterances in both languages. Although it was not specifically designed for prosodic studies, it contains enough data to get significant results. The corpus was acquired under recording studio conditions in two separate channels at 32 kHz. Speech was recorded in one of the channels and the output of a laryngograph in the other. Data were automatically labeled and manually supervised. Labeling included silences, allophonic transcription, and allophonic boundaries. This information was increased by the additional syllable and word boundaries and accent positions. Table 1 summarizes the figures of this corpus.

Feature extraction in both corpora was carried out using similar features to other experiments reported in the bibliography [1]. The features concern frequency: within word F0 range (f0_range), difference between maximum and average within word F0 (f0_maxavg_diff), difference between average and minimum within word F0 (f0_minavg_diff), difference between within word F0 average and utterance average F0 (f0_avgutt_diff); energy: within word energy

	ESMA		Boston Corpus (English)												
	(Spanish)		f1a		f2b		f3a		m1b		m2b		m3b	
Input Feature	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
f0_range	48.2	26.3	39.6	39.7	56.0	43.3	42.2	42.9	26.7	30.6	24.7	30.7	28.0	27.3
f0_maxavg_diff	22.8	16.1	19.1	21.2	25.1	21.7	18.7	21.0	14.5	19.5	12.2	16.8	13.2	13.5
f0_minavg_diff	25.4	14.6	20.4	21.8	30.9	26.4	23.5	25.8	12.3	13.8	12.4	16.0	14.7	15.4
f0_avgutt_diff	-0.8	18.4	-19.0	57.5	-5.5	42.5	-21.6	62.8	-13.4	40.5	-28.0	60.2	-15.6	45.7
e_range	18.6	8.5	13.9	6.8	16.7	6.4	13.7	6.3	12.9	6.5	12.4	6.9	11.5	5.2
e_maxavg_diff	10.0	5.4	7.7	4.4	9.2	4.1	7.7	4.0	7.7	4.2	7.8	4.8	6.9	3.6
e_minavg_diff	8.6	4.1	6.2	3.1	7.6	3.4	6.0	3.1	5.2	3.0	4.7	2.8	4.6	2.2
duration	-0.9	9.3	2.5	9.8	4.2	10.6	1.4	12.0	1.0	9.9	-0.5	12.0	1.2	9.5

Table 2. Statistics of the features for the different corpora and subcorpora. Units: f0_range, f0_maxavg_diff, f0_minavg_diff and f0_avgutt_diff in Hz; e_range,e_maxavg_diff and e_minavg_diff RMSE/100, and duration is normalized*10.

range (*e_range*), difference between maximum and average within word energy (e_maxavg_range), difference between average and minimum within word energy (e_minavg_range); duration: maximum normalized vowel nucleus duration from all the vowels of the word (normalization is done for each vowel type) (duration).

The syntactic lexical POS Tagging information has shown to be useful in the improvement of the classification results (see [15, 1]). There is no obvious correspondence between POS tags used in each corpus: BURNC corpus uses the Penn Treebank tag set (labeled using the BBN tagger [9]) and ESMA uses the EAGLES tag set for Spanish (labeled using the Freeling tagger). We decided to use the classical classification that considers the different words of the utterance to have the role of *function word* versus *content word*. This classification is broadly used for modeling Spanish intonation in Text to Speech contexts [4]. The Penn Treebank tags have been collapsed so that the *Function Words* were: EX (existential there), RP (particle), CC (coordinating conjunction), DT (determiner), IN (preposition, conjunction subordinating), WDT (Wh-determiner, TO (to preposition) and CD (cardinal number). The rest of the types of words are considered as *Content Words*. The words of the ESMA corpus are a priori classified in terms of function vs. content word as the corpus is segmented into stress groups (an stress group is formed by one content word plus the preceding function words).

Regarding to context, we focus on local effects (at the level of word and/or syllable) as the context can be highly dependent on the language and the modeling of its correspondence is beyond the scope of this paper.

2.2 The classifiers

The Weka machine learning toolkit [8] was used to build C4.5 decision trees (J48 in Weka). Different values for the confidence threshold for pruning have been tested, although the best results are obtained with the default value (0.25). The minimum number of instances per leaf is also set to the default value (2).

A Multilayer Perceptron (MLP) with a non-linear sigmoid unit is trained for each classification problem, using the Error Backpropagation learning algorithm.

	ESMA-UPC	BURNC	BURNC.f1a	BURNC.f2b	BURNC.f3a	BURNC.m1b	BURNC.m2b	BURNC.m3b
ESMA-UPC	**86.6/81.0**	72.7/76.5	75.6/76.0	74.7/76.1	76.5/77.9	82.7/76.0	73.6/75.5	75.9/74.7
BURNC	81.4/60.3	**80.5/80.4**	–	–	–	–	–	–
BURNC.f1a	71.1/72.1	–	**83.2/80.3**	79.8/78.3	76.9/74.6	78.7/77.4	80.6/80.4	78.0/76.7
BURNC.f2b	81.5/65.5	–	81.5/80.0	**84.6/82.9**	78.6/74.3	79.0/72.6	81.6/74.5	79.4/75.1
BURNC.f3a	80.9/78.6	–	80.7/79.5	79.0/79.8	**82.2/80.3**	80.3/77.8	82.4/81.6	79.1/77.3
BURNC.m1b	76.6/75.9	–	77.6/77.0	78.0/76.8	76.7/75.6	**84.7/80.8**	74.7/76.9	77.8/75.0
BURNC.m2b	74.4/63.0	–	80.5/79.5	77.8/75.1	78.3/73.5	79.1/74.4	**83.0/82.3**	78.1/75.8
BURNC.m3b	69.3/75.5	–	81.5/80.8	78.4/78.9	78.1/77.2	79.9/78.6	79.9/81.0	**81.0/76.6**

Table 3. Classification rates (in percentages) using words in terms of the presence of accent. The training corpus in the rows; the testing one in the columns. In the cells (xx/yy), where xx is the classification rate obtained with the C4.5 classifier, yy with the MLP classifier.

Several network configurations were tested, achieving the best results with the following: i) single hidden layer with 12 neurons, following the Gori results [7], more hidden units than inputs were used to achieve separation surfaces between closed classes, ii) 100 training epochs, iii) two neurons in the output layer, one for each class to be classified, then the test input vector is assigned to the class corresponding to the largest output.

Due to the different scale of the features among the training corpora, we tested different normalization techniques: the Z-Norm, Min-Max, divide by maximum and euclidean norm 1. The normalization has been processed by corpus and by speakers using the Z-Norm technique. In [6] the negative impact of imbalanced data on final result is shown. Therefore, re-sampling methods were applied: minority class example repetition [14] for the MLP classifier and Synthetic Minority Oversampling TEchnique (SMOTE) method [3] for the C4.5 classifier.

3 Results

Table 2 reports on the mean values and standard deviations of the acoustic input features of the different corpora and sub-corpora analyzed in this work. For $F0$ related variables, the differences between male and female speakers are clearly observed (μ values of $f0\_range$ go from 24.7Hz to 28.0Hz for male speakers, but they go from 42.2Hz to 56.0Hz for female speakers). F0 values seem to be more stable in the ESMA corpus (σ values goes from 14.6Hz to 26.3Hz) than in the BURNC subcorpora (σ from 13.5Hz to 62.8Hz). In the case of variables related to *energy*, there are also significant differences among the corpora. The BURNC seems to be more stable with σ going from 2.2 to 6.9 RMSE/100, versus the variability observed in the ESMA corpus, going from 4.1 to 8.5 RMSE/100. The *duration* variable shows significant differences among the diverse corpora.

Table 3 shows the classification rates that are achieved when the different corpora interchange its training and testing role. In the conventional scenarios (same corpus for training and testing; diagonal of Table 3). The results go from 80.5 to 86.6%, which are the expected ones according to the state of the art: [12] reports state of the art up-to-date results from 75.0% to 87.7% using the Boston

ESMA-UPC		BURNC.f2b		BURNC.m3b	
Feature	IG	Feature	IG	Feature	IG
f0_minavg_diff	0.18888	f0_minavg_diff	0.232	f0_minavg_diff	0.245
f0_range	0.18246	f0_range	0.214	f0_range	0.232
pos	0.17347	pos	0.199	f0_maxavg_diff	0.206
f0_avgutt_diff	0.15215	duration	0.177	e_range	0.169
f0_maxavg_diff	0.10891	f0_maxavg_diff	0.156	e_maxavg_diff	0.165
e_range	0.09695	e_range	0.152	pos	0.164
e_minavg_diff	0.08156	e_maxavg_diff	0.13	e_minavg_diff	0.15
e_maxavg_diff	0.07681	f0_avgutt_diff	0.12	duration	0.139
duration	0.0063	e_minavg_diff	0.105	f0_avgutt_diff	0.117

Table 4. Info Gain (IG), computed with the WEKA software, of the features when they are used to classify the accents in the different corpora.

Radio Corpus with words as the basic reference unit. In the cross-lingual and cross-speaker scenarios (cells out of the diagonal in the Table 3), the classification rates decrease and they are highly dependent on the sub-corpora used. The best and worst results are 82.7% and 69.3% in the cross-lingual scenario and 82.4% and 74.7% in the cross-speaker scenario. All these percentages refer to the use of decision trees that seem to be more effective than neural networks.

Table 4 compares the *Information Gain* of the different features, providing a measure of the potential loss of entropy which would be generated if the splitting of the training set was carried out in terms of the present feature [16]. The tagging of the Spanish corpus seems to rely mainly on F0 features, as the four most relevant features are related with F0 (except the *pos* feature) and the difference with respect to the energy and duration features is important. The tagging of the English corpus also seems to rely mainly on F0 features ($f0\_minavg\_diff$ and $f0\_range$ share the top ranking position in both corpora). Nevertheless, energy and duration seem to be more relevant for the English transcribers than for the Spanish ones. This behaviour seems to be dependent on the speakers: $m3b$ gives more importance to *energy* than $f3a$. The speakers *f3a* and *m3b* have been selected as they seem to be, respectively, the best and the worst for predicting the Spanish accents with the C4.5 decision tree as Table 3 reports. The feature *pos* appears as one of the most informative features in all the cases (in the BURNC.m3b corpus the feature is down in the ranking but it has a high IG value).

4 Discussion

In spite that the input features are relative magnitudes, significant differences appear between the diverse corpora (see Table 2), affecting both μ and σ. These differences were expected, independently of the cross-lingual effect, as the different recording conditions have a clear impact on the values of the input magnitudes. Thus, for example, the ESMA F0 values have been collected with a laringograph device and BURNC F0 values with a pitch tracking algorithm leading to less stable values.

The second point for discussion arising from Table 2 is that, at the time that differences between the Spanish corpus and the English one are clear, the differences between the diverse English sub-corpora are also important. The normalization of the input is thus a need in this work not only for reducing the differences that have its origin in the recording and processing conditions, but also for doing the cross-lingual comparison feasible. In [5], we present results that contrast the classification rates when the input is normalized and when not (more than ten points of accuracy can be lost in the cross-lingual scenarios).

Satisfactory classification rates seem to be obtained as reported in the previous section. The cross-lingual scenarios show lower identification rates than those achieved in the single speaker/language scenarios. Nevertheless, this decrease is comparable to the one obtained in cross-speaker scenarios in spite that the speakers were retrieved from the same corpus with similar recording conditions and with the same spoken language.

The difference between inter-speaker classification rates has its origin in the different role of the input features for characterizing the accents. This role is dependent on the speaker as the Table 4 shows, so that the different speakers seems to use the input features differently when producing the accents. The more similar is the role of the input features between two given speakers the higher the recognition rates are. This fact seems to be as relevant as the language in which the utterance has been produced.

In [5] we analyze the most common confusions, that is, situations where the classifier make a mistake by setting the wrong label to a given word. This analysis was performed by comparing the predictions of the classifiers with the labels assigned by a team of ToBI manual labelers [11]. The result is interesting because we found that cross-lingual classifiers and mono-lingual classifiers share the most common confusions. The most common mistake is to classify as *unaccented* the L* tone which represents more than 35% of all the disagreements in both cases. Furthermore, the four most common disagreements, representing more than 80% of the total amount of disagreements, are shared by both classifiers. Again, the four most common agreements representing more than the 80% of the agreements are also the same for both classifiers. This result evidences a similar behavior of the classifiers, and encourages for using cross-lingual labeling of prosodic event in combination with a posterior supervised revision of the results by human labelers in future works.

5 Conclusions and Future Work

This cross-lingual English-Spanish experiment allows to obtain promising results both in quantitative and qualitative terms. Relative high identification rates are achieved, while the confusions are consistent with the expectations according to the different shape of the Spanish ToBI accent tones. The introduction of speaker adaptation techniques, more representative input features and language dependent information added to the normalization process are expected to improve results in future work. We are currently working on the inclusion of other more expressive input features, such as Bézier interpolation parameters and the Tilt and Fujisaki parameters to improve results [6]. Furthermore, the inclusion of expert fusion techniques is also being explored to improve the classification results, as predictions of the two classifiers can be complementary in some cases.

References

1. Ananthakrishnan, S., Narayanan, S.: Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. Audio, Speech, and Language Processing, IEEE Transactions on 16(1), 216–228 (January 2008)
2. Bonafonte, A., Moreno, A.: Documentation of the upc-esma spanish database. Tech. rep., TALP Research Center, Universitat Politecnica de Catalunya, Barcelona, Spain (2008)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
4. Escudero, D., Cardeoso, V.: Applying data mining techniques to corpus based prosodic modeling speech. Speech Communication 49, 213–229 (2007)
5. Escudero, D., Gonzalez, C., Vivaracho, C., Cardenoso, V., Aguilar, L.: Analysis of inconsistencies in automatic tonal ToBI labeling. In: Proceedings of TSD 2011 (Internatinal Conference on Text, Speech and Dialogue) (in press) (2011)
6. Gonzalez, C., Vivaracho, C., Escudero, D., Cardenoso, V.: On the Automatic ToBI Accent Type Identification from Data. In: Interspeech 2010 (2010)
7. Gori, M.: Are multilayer perceptrons adequate for pattern recognition and verification? IEEE Trans. on Pattern Analysis and Machine Intelligence 20(11), 1121–1132 (November 1998)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)
9. Meteer, M., Schwartz, R.M., Weischedel, R.M.: Post: Using probabilities in language processing. In: IJCAI. pp. 960–965 (1991)
10. Ostendorf, M., Price, P., Shattuck, S.: The boston university radio news corpus. Tech. rep., Boston University (1995)
11. Prieto, P., Rosedano, P.: Transcription of Intonation of the Spanish Language. LINCOM Studies in Phonetics 06 (2010)
12. Rangarajan Sridhar, V., Bangalore, S., Narayanan, S.: Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework. IEEE Transactions on Audio, Speech, and Language Processing 16(4), 797–811 (May 2008)
13. Syrdal, A.K., Hirshberg, J., McGory, J., Beckman, M.: Automatic ToBI prediction and alignment to speed manual labeling of prosody. Speech Communication (33), 135–151 (2001)
14. Vivaracho-Pascual, Simon-Hurtado, A.: Improving ann performance for imbalanced data sets by means of the ntil technique. In: IEEE International Joint Conference on Neural Networks (18-23 July 2010)
15. Wightman, C., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing 2(4), 469–481 (October 1994)
16. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (1999)