Cross-lingual ToBI accent tones identification: preliminary results

David Escudero-Mancebo, César González-Ferreras, Carlos Vivaracho-Pascual, Valentín Cardeñoso-Payo

ECA-SIMM Laboratory, University of Valladolid, Spain
{descuder,cesargf,cevp,valen}@infor.uva.es

Abstract

This paper tackles the problem of corpus based ToBI symbol automatic identification. We focus on the binary decision of accent vs. no accent tone presence. We test a cross-lingual alternative to identify accents in a given language using supervised data learning tools trained with data of a different language. A multilayer perceptron and a C4.5 decision tree have been trained with the English Boston Radio News corpus and we test its capabilities on predicting accents in a Spanish corpus. Results are promising leading us to discuss on the application of previous work on ToBI accents multiclass classification.

Index Terms: prosody, ToBI, crosslingual, automatic recognition of prosody

1. Introduction

ToBI is a standard for representing and labelling prosodic events including tones (accent tones and boundary tones) and breaks. The tones level is used to mark the occurrence of phonological tones at appropriate points in the F0 contour. The break level is used to mark break indices, which are numbers representing the strength of the boundary between two orthographic words. The tones codification is based on the combination of two single symbols: H (high) and L (low). One of the most important prosodic features is prominence: a word or part of a word made prominent is perceived as standing out from its environment [1]. This paper focuses on this particular aspect of prosody that is also marked in the ToBI prosodic representation model[2].

ToBI has been implemented for several languages including English, German and Japanese. Despite the intesive research activity for Iberian languages; the need of a reference corpus similar as the ones existing for other languages (e.g. the Boston Radio Corpus for English [3]) is still a need both for Catalan and Spanish. The activity presented in this paper is included in the Glissando project [1], that has the aim to record and label with ToBI marks a bilingual Spanish and Catalan corpus that contains Radio news recordings and spontaneous dialogs.

Labelling a corpus with ToBI tags is an expensive procedure. In [4] it is estimated that the ToBI labelling commonly takes from 100-200 times real time. To speed up the process, automatic or semiautomatic methods seem to be a productive resource. [5] or [6] are good examples of the state of art on automatic labelling of ToBI events. For Catalan [7] presents a procedure to label break indices reducing the set of break indices merging together some of them with the aim to increase the identification results. This merging strategy is common in other studies such the ones already mentioned of [5] or [6] that combine the different type of accent tones transforming the labelling problem into a binary one to decide whether an accent is present or not.

[1]Partially funded by the Ministerio de Ciencia e Innovacion, Spanish Government Glissando project FFI2008-04982-C003-02

Here we explore a cross-lingual approach where a given corpus with ToBI labels will be used to predict the labels of a different corpus in a different language. Despite the ToBI sequences are highly dependent on the language, they codify universal functions of prosody, one of them the prominence. Thus we use the Boston Radio Corpus to train prosodic models that are used then to identify the prominence in a Spanish corpus. This cross-lingual approach is pertinent as the number of linguistic resources with ToBI labels is sparse and the number of languages that lack of this information is large.

In [8] we point out data sparseness, the high inter-symbols similarity and the large number of prosodic features potentially affecting prosodic profiles as the main difficulties for ToBI labelling automatic approximations. Here we add the normalization of the prosodic features as the challenging problem to cope with.

First we present the experimental procedure and then we present the results on crosslingual accent identification. Discussion of the future work to extend the approach to different accent type is then presented.

2. Processing of the corpus

We used the Boston University Radio News Corpus [3]. This corpus includes labels separating phonemes, syllables and words. Accents are marked with a ToBI label and a position. We take into account the 7 more frequent types of accent tones: H*, L+H*, !H*, H+!H*, L+!H*, L*, and L*+H discarding other undetermined marks like * or *?. Inspired in previous works [9, 5] we aligned the accent tones with respect to the prominent syllable and to the word that contains 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 (f1a, f2b, f3a, m1b, m2b and m3b) have been used, as shown in table 1.

The Spanish corpus used in this paper is ESMA-UPC. It was designed aiming the construction of a unit concatenative TTS system for Catalan and Spanish at the UPC (http://www.gps-tsc.upc.es) [10]. It contains three hours recordings 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 labelled and manually supervised. Labelling included silences, allophonic transcription, and allophonic boundaries. This information was increased by the additional syllable and word boundaries and stress positions. Pitch was estimated by means of glottal pulses closing time points. It eases the automatic segmentation of stress groups and the selection of the corresponding F0 profiles. Figure 2 resumes the figures of this corpus.

Similar features to other experiments reported in the bib-

	word	syllable
# utterances	421	421
H*	7587	8098
L+H*	2383	2501
!H*	2144	2358
H+!H*	586	654
L+!H*	638	666
L*	517	548
L*+H	44	48
none	13868	32450
Total	27767	47323

Table 1: Accent events in the Boston Corpus

# utterances	421
# accent groups	
Accent	7587
No accent	2383
Total	9970

Table 2: Accent events in the UPC-ESMA Corpus

liography [5] have been used. They concern to frequency: within word F0 range, difference between maximum and average within word F0, difference between average and minimum within word F0, difference between within word F0 average and utterance average F0; to energy: within word energy range, difference between maximum and average within word energy, difference between average and minimum within word energy; to duration: maximum normalized vowel nucleus duration from all the vowels of the word (normalization is done for each vowel type); and to grammatical information POS: part of speech.

3. Experimental procedure

3.1. Experimental strategy

We used two different classifiers, a C4.5 Decision Tree (DT) and a Multilayer Perceptron (MLP) Neural Network (NN), applying stratified 10-fold cross-validation. Details on the classifiers are depicted in section 3.3

First, the Accent vs No Accent classification problem (the most classical one in the literature) was approached. The goal is to contrast our systems with the state of the art. Next the more complex multiclass accent type classification problem was approached.

Once shown the trouble of the multiclass problem (high error rates in accent recognition) we focused on the data analysis, previous to continue with the classification problem. A contrast in pair of accent types was performed by applying the classifier to the easier task of binary classifications for every pair of accents. The goal is to identify similar classes as a source of confusion in the multiclass problem. Multidimensional scaling [11] is used to display these inter-class potential similarities.

3.2. Data preprocessing

Some classifiers can not handle qualitative features as the POS ones. We transformed them into quantitative characteristics by using two approaches: binary masks (one bit per POS type); and codification of the 33 values using 6 bits.

Due to the different range of the features, we applied different normalization techniques: the Z-Norm, Min-Max, divide by maximum and euclidean norm 1.

The approaches proposed for dealing with the imbalanced data can be divided into internal and external ones, i.e., at algorithmic and data level, respectively [12]. In the first, new algorithms or modifications of existing ones are proposed. In the second, the data sets are re-sampled **over-sampling** the minority class or **under-sampling** the majority class. Both options can be accomplished randomly or directed. We are interested in general solutions, so only external solutions have been applied, more specifically, re-sampling method based on minority class example repetition has been performed.

3.3. The classifiers

The Weka machine learning toolkit [13] 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). This classifier was trained with un-normalized data and qualitative POS feature.

A Multilayer Perceptron (MLP) is trained per each classification problem, using the Error Backpropagation learning algorithm. Non-linear sigmoid units are used in the hidden and output layers because they showed better performance than $tanh$ ones in our experiments. Several network configurations were tested to define the final MLP configuration: i) single hidden layer, ii) training epochs equal to 100, iii) although Gori [14] has demonstrated that only using more hidden units than inputs the separation surfaces between classes in the pattern space can be closed, the results showed that using more than 16 hidden units is not worth it, iv) as many units as classes are used in the output layer, one per each class to classify.

To train the MLP unsaturated desired outputs [15] were tested. The chosen ones, however, were 1.0 for the output corresponding to the training vector class and 0.0 for the rest, since a better performance was achieved.

Although the assumptions to approximate the MLP output to a posteriori probability are not fulfilled [15], given a test vector x_i, each output of the MLP, trained to distinguish between n classes C_j, can be seen as the estimation of the membership degree, $\Gamma(C_j/x_i)$, of vector x_i to class C_j. Then, the input vector is assigned, in accordance with this probabilistic output interpretation, as follow: $x_i \in C_j$ with $j = \arg\max_j \Gamma(C_j/x_i)$. If all the outputs have the same value, that is very rare, the input is assigned to the most probable class, i.e., the largest.

The codification alternative showed better performance to transform the POS feature (besides, the input vector is smaller). Z-Norm was the chosen to normalize the feature ranges, since it showed the best performance.

4. Results

When the classifiers are applied to the *Accent* vs *No Accent* binary decision, results are close to the expected according to the state of the art: we achieved 84.7% with NN and 82.7% with DT. [6] summaries the state of the art up to date reporting results from 75.0% to 87.7%. These results have been obtained using the Boston Radio Corpus using all the input features mentioned in the last paragraph of section 2.

When we change to cross-lingual scenarios, first limitation is that not all the prosodic features are immediate to be used. For example POS Tags can be highly dependent on the language. We decided in this preliminary approach to make use of the F0

```
f0_minavg_diff <= 17.9909
|   f0_range <= 8.9: none (4224.0/373.0)
|   f0_range > 8.9
|   |   e_maxavg_diff <= 792.058: none (6008.0/1616.0)
|   |   e_maxavg_diff > 792.058
|   |   |   e_maxavg_diff <= 1054.15: none (2186.0/924.0)
|   |   |   e_maxavg_diff > 1054.15: accent (1916.0/874.0)
f0_minavg_diff > 17.9909
|   f0_avgutt_diff <= -28.3332
|   |   f0_minavg_diff <= 52.6954: none (1297.0/389.0)
|   |   f0_minavg_diff > 52.6954
|   |   |   e_maxavg_diff <= 965.426
|   |   |   |   f0_avgutt_diff <= -35.3209: none (166.0/64.0)
|   |   |   |   f0_avgutt_diff > -35.3209
|   |   |   |   |   e_range <= 1123.68: none (25.0/9.0)
|   |   |   |   |   e_range > 1123.68: accent (85.0/23.0)
|   |   |   e_maxavg_diff > 965.426: accent (272.0/52.0)
|   f0_avgutt_diff > -28.3332: accent (11588.0/2388.0)
```

Figure 1: Decision tree C4.5. Simplified version with pruning confidence of 0.001 (default is 0.25)

Classified as –>	Accent	No Accent
Accent	1698	634
No Accent	616	1698

Table A

Classified as –>	Accent	No Accent
Accent	1663	669
No Accent	564	1750

Table B

Table 3: Classification results using the C4.5 decision tree. Table A uses input features of frequency and energy. Table B only uses F0 features.

and energy features.

Second difference with respect to the mono-lingual scenario is the need to normalize the input features. Figure 1 shows the simplified decision tree resulting when the algorithm is feed with not normalized data. Despite the features are relative only the F0 features would support a comparison between corpora and speakers. Energy features are still highly dependent, not only on the speaker, but also on recording conditions. A change on the energy scale between corpora would affect dramatically results. The data in Boston Radio Corpus and in ESMA-UPC have been normalized separately using the same z-norm.

Table 3 shows the confusion matrix using normalized fundamental frequency features and energy when the decision tree is trained with the Boston Radio Corpus and tested with the ESMA-UPC corpus. Despite the most relevant set of features seems to be the one relating to F0, energy seems to have a rol to take into account.

Table 4 compares recognition rates in pairs of corpus.

Modelling-Testing corpora	Accent	No Accent
Boston-Boston	73.4%	76.7%
ESMA-ESMA	80.6%	72.1%
Boston-ESMA	75.6%	71.6%
ESMA-Boston	75.1%	73.4%

Table 4: MLP Neural Network classification rates results. Input features correspond to F0 and energy features.

When the pair modelling-testing correspond to samples of the same corpus, results are better. Nevertheless, the figures corresponding to cross-lingual experiments are not far away from the mono-lingual ones and are encouraging to make use of other features relating to duration or to grammar to increase the identification results.

No Accent classification seems to be more difficult in the ESMA corpus than in the Boston one (72.1% and 71.6% versus 76.7% and 73.4%). This is because the ESMA corpus is divided into stress groups not in words. The first word (in any) of the stress group is un-stressed by definition, so that there is an important number of words that would belong to the No Accent category that would be identified easily by the classifier.

5. Discussion and future work

In [8] we show the importance of the proper selection of the input feature in order to improve result by entering a more expressive parameterization technique based on the use of Bzier functions [16]. When we perform multiclass classification with the data of the Boston Radio Corpus results dramatically decrease (see table 6). Nevertheless, the table 5 shows the classification rates for every pair of classes. The use of the Bézier coefficients outperforms the results in both classifiers. Although in the *Accent* vs *No Accent* the improvement is very low, in the multiclass and in the pairwise classification problem the use of Bézier coefficients permits to improve results. For example !H* increases its rates from 18.7 to 29.3 in multiclass classification, and it also increases its performance with respect to all the other classes in the pairwise classification problem. In this paper we made use of a set of features getting inspiration from other state of the art studies, but a deeper analysis and the test of alternative features seems to be a need.

The significant differences between the two type of classifiers opens the door to an alternative research such as the use of expert fusion. By combining results of different classifiers or by specialising experts in different type of accents we expect to improve the performance.

In this work we focus on the binary accent vs. no accent problem. The questions arising at this moment is weather is possible or not to extend this approach for the recognition of different ToBI accent. The immediate answer is no, and reason is that ToBI sequences are highly language dependent. Furthermore, ToBI identification results in the corpus are still very poor. Finally, the high level of inconsistency in labellers tagging for the case of Sp-ToBI[17] does not encourage to explore this possibility.

6. Conclusions

We have presented a cross-lingual experience on ToBI accent identification. The two corpora used have been presented and the experimental strategy has been described. Results indicate that this is a promising alternative to analyze in deeper detail in future work.

Related work [8] has shown us the difficulties of doing multiclass ToBI accent classification, but we have also learn the future work to be done to cope with data sparseness: using more expressive prosodic features and using more powerful learning tools and strategies.

7. References

[1] J. Terken, “Fundamental frequency and perceived prominence,” *Journal of Acoustics of America*, vol. 89, no. 4, pp. 1768–1776,

(a) MLP Neural Networks

	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*		60,7	59,8	77,8	66,7	85,8	98,6	86,8
L+H*	59,0		71,0	77,7	64,6	83,4	96,7	86,6
!H*	65,4	68,4		71,7	59,5	77,8	96,5	85,5
H+!H*	61,4	73,4	60,2		67,5	66,1	92,4	69,5
L+!H*	51,9	53,0	53,3	78,6		80,3	90,3	71,9
L*	73,3	79,0	66,7	64,6	78,5		91,5	68,3
L*+H	4,0	6,0	12,0	18,0	16,0	32,0		4,0
none	81,1	87,9	82,7	81,9	89,9	85,2	99,4	

	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*		67,1	64,8	84,2	72,1	93,4	99,0	85,0
L+H*	60,8		72,8	85,5	65,8	92,8	97,8	87,3
!H*	65,5	74,2		77,9	65,9	86,7	97,8	84,0
H+!H*	62,2	72,4	61,4		74,9	78,8	96,8	63,4
L+!H*	48,3	52,0	56,4	78,0		88,8	90,8	73,3
L*	71,9	78,3	73,7	68,1	89,8		92,9	63,7
L*+H	8,0	12,0	20,0	56,0	38,0	36,0		26,0
none	85,6	90,2	84,2	85,7	94,0	90,4	99,6	

(b) C4.5 Decision Trees

	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*		71,3	68,3	91,8	89,9	93,6	98,3	80,6
L+H*	41,5		67,4	83,7	73,5	88,5	97,2	69,6
!H*	50,8	62,6		79,2	66,6	82,6	97,1	59,8
H+!H*	29,7	54,6	43,2		71,5	60,6	90,3	22,7
L+!H*	14,7	33,7	42,3	74,1		77,3	91,5	47,8
L*	33,8	65,8	49,9	66,2	77,0		91,7	21,7
L*+H	6,8	11,4	9,1	25,0	22,7	38,6		9,1
none	82,2	92,2	90,5	95,8	96,2	96,4	98,7	

	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*		75,6	75,3	92,8	90,4	95,3	98,2	78,1
L+H*	36,5		71,2	86,4	77,1	93,0	97,0	70,0
!H*	43,6	68,6		81,3	77,1	87,8	97,2	57,6
H+!H*	33,1	63,5	47,3		74,4	71,7	93,5	25,3
L+!H*	16,8	32,6	38,9	74,1		86,8	93,3	48,7
L*	59,2	81,2	70,4	70,8	85,9		91,1	29,2
L*+H	2,3	13,6	15,9	29,5	34,1	43,2		15,9
none	84,2	93,4	91,7	95,6	97,0	96,6	99,0	

Without Bézier parameters With Bézier parameters

Table 5: Accuracy (in %) of the pairwise classifiers using neural networks (a) and decision trees (b). In both cases, individual class success rate is shown. Tables on the left show results without Bézier coefficients and the ones on the right with Bézier coefficients. Position i, j of the table represents the success rate of the class i in the classifier $i\ vs.\ j$.

	C4.5 DT		**MLP NN**	
Acc Type	**NBez**	**Bez**	**NBez**	**Bez**
H*	44.4	45.5	21.5	22.1
L+H*	22.7	25.6	35.4	41.0
!H*	18.1	21.9	18.7	29.3
H+!H*	9.4	12.5	32.7	42.7
L+!H*	6.6	7.1	28.8	31.1
L*	11.4	17.6	43.5	59.6
L*+H	0.0	2.3	0.0	2.0
none	75.3	75.5	68.3	68.2
Acc-NoAcc	82.6	82.7	83.0	84.7

Table 6: Accuracy (in %) of the Decision Trees (column *C4.5 DT*) and Neural Networks (column *MLP NN*) in the multiclass accent type and accent vs. no accent (last row) recognition tasks, when the Béziers coefficients are used (column *Bez*) and not used (column *NBez*).

1991.

[2] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labelling English prosody," in *Proceedings of ICSLP-1992*, 1992, pp. 867–870.

[3] M. Ostendorf, P. Price, and S. Shattuck, "The boston university radio news corpus," Boston University, Tech. Rep., 1995.

[4] A. K. Syrdal, J. Hirshberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speech manual labeling of prosody," *Speech Communication*, no. 33, pp. 135–151, 2001.

[5] S. Ananthakrishnan and S. Narayanan, "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 216–228, January 2008.

[6] V. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 797–811, May 2008.

[7] L. Aguilar, A. Bonafonte, F. Campillo, and D. Escudero, "Determining Intonational Boundaries from the Acoustic Signal," in *Proceedings of Interspeech 2009*, 2009, pp. 2447–2450.

[8] C. Gonzalez, C. Vivaracho, D. Escudero, and V. Cardenoso, "On the automatic ToBI accent type identification from data," in *Proceedings of Interspeech 2010*, 2010.

[9] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, October 1994.

[10] A. Ferrer, "Sintesi de la parla per concatenaci basada en la selecci," Ph.D. dissertation, Dpto. de Teora del Senyal i Comunicacions, Universidad Politcnica de Catalua, Espaa, 2001.

[11] R. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley and Sons, 2001.

[12] A. Vivaracho-Pascual, Simon-Hurtado, "Improving ann performance for imbalanced data sets by means of the ntil technique," in *Accepted to the IEEE International Joint Conference on Neural Networks*, 18-23 July 2010.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[14] M. Gori, "Are multilayer perceptrons adequate for pattern recognition and verification?" *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1121–1132, November 1998.

[15] S. Lawrence, I. Burns, A. Back, A. Chung Tsoi, and C. L. Giles, "Neural networks classification and prior class probabilities," *Lecture Notes in Computer Science State-of-the-Art Surveys*, pp. 299–314, 1998.

[16] D. Escudero and V. C. A. Bonafonte, "Corpus based extraction of quantitative prosodic parameters of stress groups in spanish," in *Proceedings of ICASSP 2002*, vol. 1, 2002, pp. 481–484.

[17] D. Escudero and L. Aguilar, "Procedure for assessing the reliability of prosodic judgements using Sp-TOBI labeling system," in *Proceedings of Prosody 2010*, 2010.