On the Automatic ToBI Accent Type Identification from Data

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

Department of Computer Science, University of Valladolid, Spain

{cesargf,cevp,descuder,valen}@infor.uva.es

Abstract

This contribution faces the ToBI accent recognition problem with the goal of multiclass identification vs. the more conservative *Accent* vs. *No Accent* approach. A neural network and a decision tree are used for automatic recognition of the ToBI accents in the Boston Radio Corpus. Multiclass classification results show the difficulty of the problem and the impact of imbalanced classes. A study of the confusion/similarity between accent types, based on in-pair recognition rates, shows its impact on the overall performance. More expressive F0 contours parametrization techniques have been used to improve recognition rates.

Index Terms: prosody, ToBI, automatic recognition

1. Introduction

The detection of prosodic events have been tackled from different perspective in the state of the art, since the pioneer references in [1] to the more recent ones in [2, 3]. On what concerns to the ToBI Accent identification, state of the art approaches mainly focus on the binary Accent vs No Accent decision with exceptions like [4] clustering 9 ToBI tags into 5 classes. The mentioned references point out the difficulty of the multi accent classification task but they do not offer results on this more ambitious challenge. In this work we investigate the sources for these difficulties, trying to deal with them, reporting preliminary results on this ongoing research.

Increasing of the granularity of the Accent vs. No Accent decision is important because the ToBI accent sequences can be associated to a meaning or to a prosodic function projected in the prosodic acoustic characteristic of the utterance. Consequently the identification of such ToBI accent sequences from the utterances can have applications for the detection of speech acts, disambiguation or speech recognition. Furthermore the modeling of the relationship between such sequences of accents and the corresponding prosodic shape would be very useful to increase the naturalness in text to speech applications. In the context of the Glissando project[1], in which this work is performed, we expect to increase the speed and performance of manual ToBI tagging corpora offering an automatic proposal for manual labelers to revise (inspired in [5]).

We focus on three reasons making difficult the ToBI accent identification task. First is the presence of imbalanced classes in sparse training corpora, second is the intrinsic similarity of some of the ToBI accents pairs and third is the selection of relevant classification input features. Concerning to the first of the reasons, the state of the art offers very few ToBI tagged corpora, Boston Radio News one is probably the main reference. This corpus is clearly imbalanced in what concerns to the presence of different ToBI accent type (for example H* class has more than ten times samples than L* class). In the Glissando project we have produced a radio news Spanish and Catalan corpus selecting radio news stories to ensure a minimum number of a-priori prosodic classes, but the class cardinality contrast still remains [6]. Independently of the corpus, it is clear that these differences depend on the language so that it is a fact to assume. The problem is that some classifiers are very sensitive to these situations getting specialized on the recognition of the more populated classes. In this paper we show the negative effect of this fact by comparing results using C4.5 decision tree classifier and a Multilayer Perceptron classifier including specific techniques to reduce the impact of imbalanced input data.

Concerning to the intrinsic similarities of ToBI accents, despite of the inter-transcriber agreement quality control procedures to be applied, it is undeniable that some of the ToBI accents are easy to confuse. [7] reports on ToBI labeling inconsistencies based on empirical intertranscriber judgments and on the opinions of the labelers about its conceptual similarity. Even the Boston Radio Corpus, with intertranscriber agreements rates higher than 90% [8], also reflects this fact including comments of the labelers with doubts about the judgments. We hypothesize that these difficulties on the label assignment can be projected into an objective interclass similarity affecting the confusion of the automatic labeling assignment.

The set of features to feed the classifier is determinant on its final results. In the Accent vs No Accent problem of the state of the art, the classifiers use statistics of F0, energy and duration in the prosodic unit to study (syllable or word) and also pseudolinguistic features. We hypothesize that the type of accent discrimination needs to incorporate other more expressive prosodic features representing the temporal evolution of the F0 pattern in the intonation unit. In this paper we analyze the impact of including quantitative F0 contour approximation parameters [9] that have shown to be useful on modeling intonation [10] in text-to-speech applications.

First the processing of the corpus and the experimental procedure are presented. Next results and future work are reported and discussed.

2. Processing of the corpus

We used the Boston University Radio News Corpus [8]. 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 [1, 2] we aligned the accent tones with respect to the prominent syllable and to the word that contains it (words with more than

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

	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.

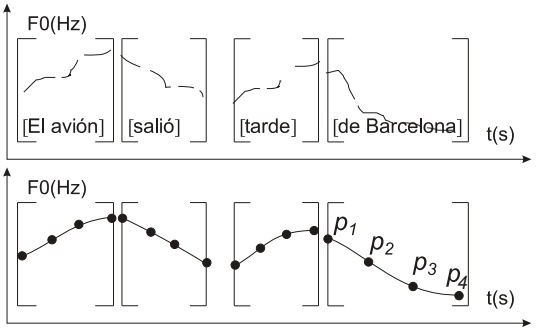

Figure 1: Example of Bézier function fitting stylization (from [10]). A set of quantitative parameters p_0, p_1, p_2, p_3 represent the temporal evolution of the F0 contour in a given intonation unit (stress group, word, syllable...).

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.

Similar features to other experiments reported in the bibliography [2] 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 pseudo-grammatical information POS: part of speech.

As one of our goals is to measure the impact of alternative prosodic features that represent the evolution of the F0 contour, we included a set of coefficients representing the fitting Bézier function that stylizes the F0 contour in the intonation units (see illustration in figure 1 and [9] for details). In this work we used four interpolation points to represent the F0 tendency in the syllables and in the words.

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

	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 2: 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*).

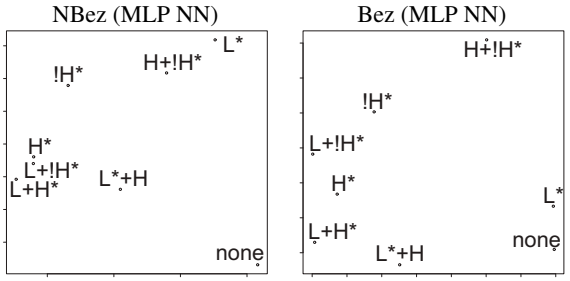

Figure 2: Inter-class distance multidimensional scaling plot. Distances between every pair of classes is proportional to the recognition accuracy showed inter-class in table 3. Scale and units are not relevant.

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 (last row of the table 2): we achieved 84.7% with NN and 82.7% with DT. [3] summaries the state of the art up to date reporting results from 75.0% to 87.7%.

When we perform multiclass classification results dramatically decrease (see table 2). Despite the *noAccent* class is still discriminated with a high rate (*Acc Type=none*), the confusion between the accents is very high. Note the differences between DT and NN. DT seems to identify properly the classes *none* and *H**. Nevertheless, these higher rates seem to be the consequence of the higher number of samples in the classes (see table 1): DT classifier get specialized on the recognition of the most populated classes to increase its overall scores. This effect is compensated in the NN classifier with better results for the rest of the classes.

The table 3 shows the classification rates for every pair of classes. The DT is still more sensitive with regard to the imbalance problem (the value of the cell i, j is usually higher than the value in the cell j, i if the number of samples in the i class is higher than the number of samples in the class j). Results with both classification techniques, DT and NN, are reasonably satisfactory: only very infrequent classes have high error rates like the class *L*+H*.

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.

We transformed the matrices of table 3 into symmetrical matrices by using the geometrical mean of the elements i, j and j, i. These distance matrices can be interpreted as inter-class similarity matrices as the higher the recognition rate, the easier the distinction between the classes. Plots in figure 2 results from the application of multidimensional scaling to the distance matrices ($cmdscale$ of R[2]). The closer the classes, the more difficult its distinction has been for the NN classifier. Note that the use of Bézier (plot on the right position) features spreads a potential cluster formed by the classes *L+!H*, L+H** and H* increasing the capabilities of the classifier to recognize them.

5. Discussion and future work

Previous work on accent type identification, reported in [4], performed a clustering of ToBI accents, resulting in five classes: H* grouping (H*, L+H*), !H* (!H*, L+!H*, H+!H*), L* (L*,L*+H) and ?* (?*) and UA for unaccented syllables. This knowledge based grouping is supported by the ToBI standard rules, but recognition results reported in this paper invite to the use of alternative pattern recognition supported grouping of classes. A hierarchical recognition scheme seems to be an alternative, where different classifiers get specialized in the separation of different clusters of ToBI Accents.

In this first approach, our interest has been to study the inter ToBI class confusion locally. The combination of a-priori judgments with a posteriori in-sequence evidences is expected to separate classes like H* or L+H* with respect to its downstepped version !H* or L+!H* that in our results appear clearly confused.

The parametrization technique of the F0 contour that we have proposed (Bézier interpolation points) has shown to be efficient to support the discrimination among different type of accent. There are other alternative parametrization techniques like Fujisaki, Tilt, IntSint (see [16] for a review) to be contrasted to show its efficiency on the characterization of the different ToBI accents.

Table 3 shows that despite the NN classifiers generally perform better, some of the classes are better discriminated with the DT classifier. Thus, for example, DT separates better H* and !H* than NN. In parallel we repeated experiments using syllables instead of words as basic unit and boundary tones instead of pitch accents obtaining similar results. We decided to focus

[2]The R Project for Statistical Computing, `http://www.r-project.org`

(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 3: 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 $i\ vs.\ j$.

on the case word and accent because the confusion inter-class is more accused (worst error rates in general). Nevertheless we observed that some classes are better discriminated using syllables, while other classes are better discriminated using words. These evidences lead us to start new works on the use of *expert fusion* to combine results of different classifiers.

6. Conclusions

In this communication we have tackled the automatic classification of ToBI accents problem. We have observed the difficulty of the task and mined into the reasons for this difficulty, concluding that the classifier to be used must be prepared to cope with scarce data with clearly imbalanced classed. In our case, MLP operates clearly better than a C4.5 decision tree classifier.

Second we have observed that some of the classes are easy to separate at the time that other pair of classes seems to be too close to be discriminated. The inclusion of more expressive prosodic features in the input of the classifier has shown to be effective to increase the classification results. Furthermore, we have identified the closest accents, candidate to be merged in future experiments with the goal to increase the classification rate.

7. References

[1] 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.

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

[3] 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.

[4] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, T. Yoon, and S. Chavarria, "Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus," *Speech Communication*, no. 46, pp. 418–439, 2005.

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

[6] D. Escudero-Mancebo, C. Gonzalez-Ferreras, J. M. Garrido, E. Rodero, L. Aguilar, and A. Bonafonte, "Combining greedy algorithms with expert guided manipulation for the definition of a balanced prosodic spanish-catalan radio news corpus," in *Prosody*, 2010.

[7] R. Herman and J. McGory, "The conceptual similarity of intonational tones and its effects on intertranscriber reliability," *Language and Speech*, vol. 45, pp. 1–36, 2002.

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

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

[10] D. Escudero and V. Cardeñoso, "Applying data mining techniques to corpus based prosodic modeling speech," *Speech Communication*, vol. 49, pp. 213–229, 2007.

[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 Transactions 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] A. Botinis, B. Granstrom, and B. Moebius, "Developments and paradigms in intonation research," *Speech Communication*, vol. 33, pp. 263–296, July 2001.