Multiclass Pitch Accent Classification for Assisting Manual Prosodic Labeling*

David Escudero-Mancebo[1], Francisco Vizcaíno-Ortega[2], César González-Ferreras[1], Carlos Vivaracho-Pascual[1], Mercedes Cabrera-Abreu[2], Eva Estebas-Vilaplana[3], and Valentín Cardeñoso-Payo[1]

[1] Department of Computer Science, Universidad de Valladolid, Valladolid, Spain
{descuder,cesargf,cevp,valen}@infor.uva.es
[2] Department of Modern Languages, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
{fvizcaino,mcabrera}@dfm.ulpgc.es
[3] Department of Modern Languages, Universidad Nacional de Educación a Distancia, Madrid, Spain
eestebas@flog.uned.es

Abstract. In this paper, we present an experiment on computer assisted prosodic labeling in which a labeler team validates or corrects ToBI pitch accents automatically predicted by a classifier. The innovative aspect of this automatic system is that it is not a deterministic prediction model, it offers the human transcriber more than one label per word (multi-class classification) and it is the transcriber who must decide on their appropriateness. The results make it evident that the procedure is indeed effective, since it not only reduces manual transcription time, but also improves global inter-transcriber consistency.

Keywords: Computer Assisted Prosodic Labeling, ToBI Labeling, Automatic Prosodic Labeling

1 Introduction

Prosodic labeling is a useful tool which provides relevant information for many applications. One such piece of information is the emphasis or accent whereby some parts of the utterance are highlighted. Knowing what parts of the utterance are accented helps improve various speech technology applications. Thus, detecting the prominent syllable in a word can be used to perform lexical disambiguation in Automatic Speech Recognition. In Dialog System, the knowledge of the emphatic words can serve to interpret or classify a message from either a semantic or a pragmatic viewpoint. As for Text-to-Speech synthesis, the correct identification between the prosodic form and the prosodic function is essential

* This work has been partially supported by Ministerio de Ciencia e Innovacion, Spanish Government (Glissando projects FFI2008-04982-C03-02 and FFI2011-29559-C02-01,2) and by Consejería de Educación de la Junta de Castilla y León (project VA322A11-2)

to give expression to the message. ToBI[2] is at present the most widely used prosodic annotation system. It distinguishes different pitch accents in terms of their linguistic function (phonological aspect) and their characteristic shape, either as monotonal accents, namely H(igh) and L(ow), or as a combination of these two (phonetic aspect).

Automatic ToBI accent type identification has been recently investigated with the application of a novel classification technique that allowed us to achieve a 70.8% identification accuracy rate in multi-class scenarios [5]. Binary decisions regarding presence versus absence of pitch accent reach 90% [1,9]. One of the reasons why it is difficult to overcome this recognition rate is the high level of uncertainty concerning the labelers' judgements of some ToBI annotations. The study reported in [8] supports this fact empirically, in which different transcribers are asked about the pairs of labels they find most confusing. Furthermore, some ToBI labeled corpora (the Boston University Radio News Corpus [10] among them) include notes of the transcribers stating that a second label could also be used for tagging a given accent. Taking such ambiguity into consideration, in this paper we present an experiment in which we adapted the classifier presented in [5] so that more than one accent type is assigned to each word. Various alternative pitch accents are offered to reproduce the uncertainty exhibited in the labelers' judgements.

Manual prosodic labeling is a costly task, requiring substantial time by the transcriber. Well-trained human labelers are needed to perform an activity whose duration has been estimated to take from 100-200 times real time [13]. Offering the human labeler automatic predictions for them to correct or validate is a useful strategy that has been successfully tested in previous work [13]. In the present paper, we show that assisting human transcribers with multiple ToBI prosodic labels assigned by our classifier allows a significant reduction in manual transcription time.

In the case of large speech corpora several labelers work in parallel on different parts of the corpus in order to be more efficient. If we also want the process to be effective, in a context where there is a high level of uncertainty in the labelers' judgements, we must ensure that they all follow the same labeling criteria. In this paper, we demonstrate that assisting human transcribers with ToBI prosodic labels predicted by our classifier also implies an improvement in consistency among them.

The structure of the paper is as follows. First, we present the experimental procedure (section 2); we then provide a detailed description of the corpus used (section 2.1), the automatic classifier (section 2.2), the labeler team and the labeling procedure (section 2.3), and the metrics used to assess the procedure (section 2.4). The results reveal both a significant reduction in manual transcription time and an improvement in consistency among transcribers (section 3). The conclusions are presented in section 4.

2 Experimental Procedure

2.1 Corpus and Parameterization

We used the Boston University Radio News Corpus [10]. 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 pitch accent tones: H*, L+H*, !H*, H+!H*, L+!H*, L*, and L*+H, discarding other undetermined marks like * or *?. Inspired in previous works [1,14] we aligned the accent tones with respect to the word. We used all the utterances in the corpus with TOBI labels, from all the speakers (f1a, f2b, f3a, m1b, m2b and m3b). The total number of samples that have been considered are: H*: 7587, L+H*: 2383, !H*: 2144, H+!H*: 586, L+!H*: 638, L*: 517, L*+H: 44 and no-accent: 13868.

Similar features to other experiments reported in the state of the art have been used [1]. They concern to frequency, energy, duration and pseudo-grammatical information POS. Furthermore, the impact of alternative prosodic features that represent the evolution of the F0 contour has been considered. We included a set of coefficients representing the fitting Bézier function that stylizes the F0 contour and Tilt parameters (see [5] for details).

2.2 Automatic Labeling

The complex multi-class classification problem is divided into several simpler ones by means of pairwise coupling. We propose to combine two-class classifiers to achieve the multi-class classification, because two-class problems provide high accuracy results. Besides, the complementarity of Artificial Neural Networks (ANN) and Decision Trees (DT) classifiers has been exploited to improve the final system, combining their outputs using fusion methods (see figure 1).

The pairwise coupled approach divides a given multiclass classification problem into binary classification sub-problems whose results must be combined to get the final classification result [7,15]. According to this approach, by $\hat{P}(l|x,\lambda^{k}_{l,m})$ we mean an estimation of the probability $P(y = l|x, \quad y = l \vee m)$, where l and m are two different prosodic labels; x is the input of the classifier (in our case the prosodic features described in the previous section); y is the class label;$\lambda^{k}_{l,m}$ is a pairwise classifier of the type k (in our case a decision tree or a neural network) trained to separate the classes l and m.

From these estimators we build $\hat{P}(l|x,\lambda^{k})$, obtained with the classifier of type k, by using fusion operation:

$$\hat{P}(l|x,\lambda^{k}) = \bigotimes_{\substack{l,m=1..C \\ l \neq m}} \hat{P}(l|x,\lambda^{k}_{l,m}) \tag{1}$$

where C is the number of classes or prosodic labels and $\bigotimes$ is the fusion operator.

We step forward to fuse the results of K independent type of classifiers so that the final estimation of $P(l|x)$ would be $\hat{P}(l|x)$ computed as:

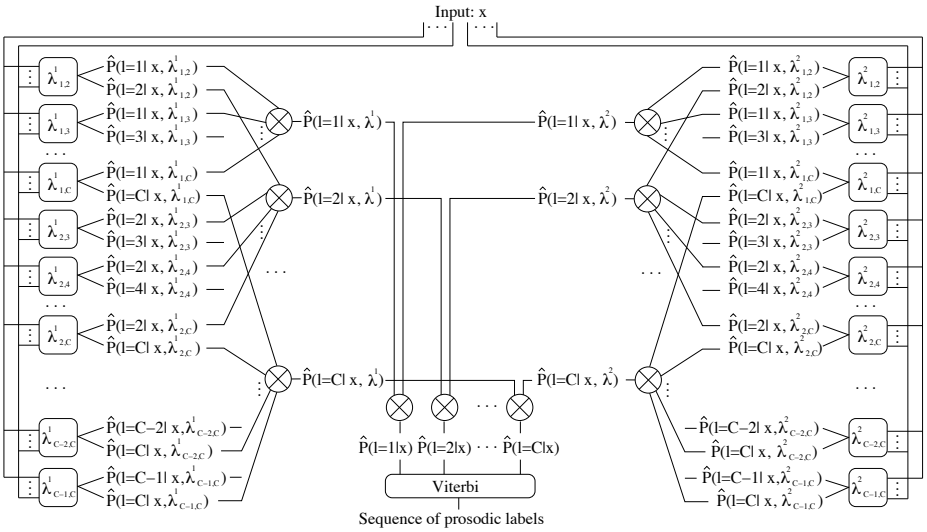

Fig. 1. Diagram of the classification procedure used in experiments.

$$\hat{P}(l|x) = \bigotimes_{k=1..K} \hat{P}(l|x,\lambda^k) \tag{2}$$

The system proposed can be seen graphically in Fig. 1. There are as many classifiers as combinations of pairs of C classes: $\frac{C\cdot(C-1)}{2}$. Each classifier, $\lambda^k_{l,m}$, provides the a posteriori probabilities estimation $\hat{P}(l|x,\lambda^k_{l,m})$ and $\hat{P}(m|x,\lambda^k_{l,m})$. The results of the classifiers are fused as described in (1) and (2). Finally, the classification rule selects the label l^* so that $l^* = \text{argmax}_l\, \hat{P}(l|x)$

In [5] we used Neural Networks as λ^1 and Decision Trees as λ^2. The fusion of results is based on the product rule and the best sequence selection is done by applying the Viterbi algorithm. This proposal, together with an adequate feature extraction, which includes the use of Tilt and Bézier parameters, allows us to achieve a total classification accuracy of 70.8% for pitch accents, 84.2% for boundary tones and 74.6% for break indices on the Boston University Radio News Corpus.

The output of the classifier is the label that obtains the highest decision score, and a number of other potential labels that are typically confused with the one selected are discarded. In this work we decided to offer the evaluation team the first three labels which were ranked as most likely for them to decide on the appropriate one. Table 1 shows how the degree of accuracy of the system increases when more than one label is taken into account. As the accuracy increases from 70.8% to 92.4%, the degree of confidence of the labelers is also expected to increase as well.

Table 1. Accuracy of pitch accent tone classification using 1 label, 2 labels and 3 labels

	1 label	2 labels	3 labels
H*	72.5%	92.1%	97.2%
L+H*	25.3%	73.4%	87.7%
!H*	35.2%	63.9%	74.9%
H+!H*	12.1%	34.6%	49.5%
L+!H*	6.0%	27.3%	45.6%
L*	11.4%	44.1%	55.9%
L*+H	0.0%	0.0%	4.5%
none	91.0%	96.5%	98.9%
Total	70.8%	86.8%	92.4%

2.3 Labeling Team and Procedure

The three transcribers who participate in this study have extensive experience with the ToBI labeling system. Indeed, one of them (T1 from now) has been involved in professional labeling projects in which she was requested to label more than one hour of a corpus for Text-To-Speech purposes.

The transcribers were asked to perform the tagging task in two different scenarios: with and without automatic prosodic labeling. Praat[4] software was used in both scenarios. In the assisted scenario, the manual labeler was confronted with TextGrid files containing five tiers: one with the orthographic transcription, three with different ToBI labels, and one tier which was empty. Transcribers had previously been informed that the labels in the tiers were ranked, being the most probable according to the automatic classifier the one in the top tier. Transcribers had to fill in the bottom tier with a number indicating which of the above tiers contained the most appropriate label (see Figure 2). If none of them seemed adequate, the transcribers supplied their own label. In the unassisted scenario, only two tiers were provided: one with the orthographic transcription and one empty tier to be filled in by the labelers. Table 3 summarizes the quantity of items that have been labeled in both scenarios.

2.4 Metrics

In order to measure the consistency of the judgments of the different labelers, we used both the *pairwise transcriber agreement* and the *kappa index*. We briefly describe these metrics in this section (see [12] for a more formal and detailed description).

The *pairwise transcriber agreement* was measured by counting the number of labeling agreements for all pairs of transcribers. That is, 4 transcribers (T1, T2, T3, T4) would produce 6 possible transcriber pairs (T1T2, T1T3, T1T4, T2T3, T2T4, T3T4). The criterion is conservative: if 3 out of 4 transcribers agree, only

[4] http://www.fon.hum.uva.nl/praat

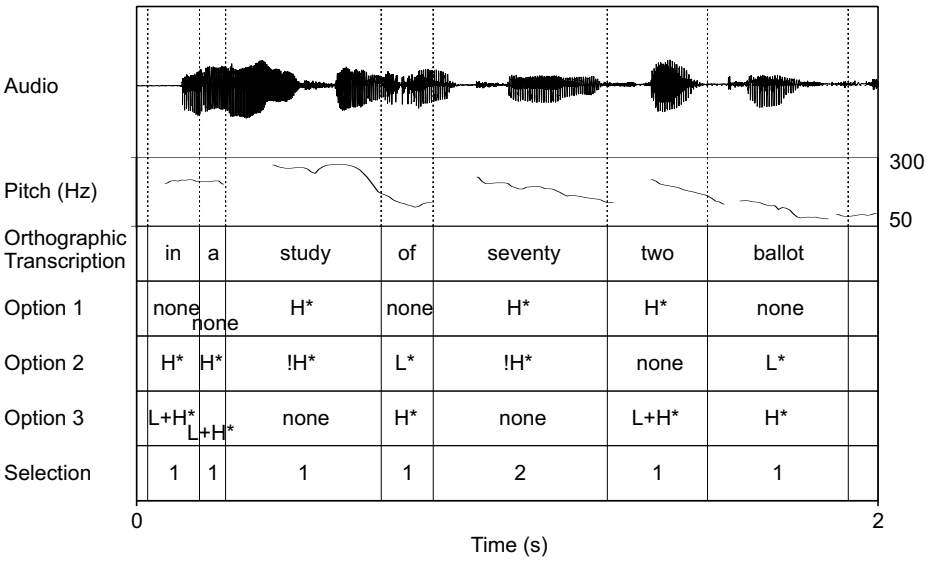

Fig. 2. Praat interface in the assisted labeling scenario.

3 out of 6 pairs will match, making the agreement rate 50% (agreement = agree / (disagree + agree)).

The Cohen's kappa, which works for two raters, and Fleiss' kappa[4], an adaptation that works for any fixed number of raters, improve upon the pairwise transcriber agreement in that they take into account the amount of agreement that could be expected to occur by chance. If a fixed number of people assign numerical ratings to a number of items then the kappa will give a measure for how consistent the ratings are. The kappa, κ, can be defined as, $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$. The factor $1-\bar{P}_e$ gives the degree of agreement that is attainable above chance, and, $\bar{P}-\bar{P}_e$ gives the degree of agreement actually achieved above chance.

3 Results and Discussion

Table 3 compares the degree of inter-transcriber consistency in both the assisted and the unassisted scenario with results from other consistency tests found in the state of the art. The global consistency rate among transcribers increases from 0.51/63.9% in the unassisted scenario to 0.55/67.0% in the assisted one. Table 2 shows that consistency increases in each pair, reaching more than 5% in the pair T1-T3.

Table 4 shows the use made by the transcribers of the different options. As can be seen, they mainly select the label corresponding to the top tier, namely, the prediction ranked first by the classifier. There are differences among the transcribers: whereas T2 and T3 use the option *Other* more frequently than T1, the latter resorts to the first option more often than T2 and T3, 71% vs. 57% and

Table 2. Inter-labeler agreement expressed as kappa index/pairwise inter-transcriber agreement. T1, T2 and T3 are the transcribers.

	T1-T2	T1-T3	T2-T3
Un-assisted	0.44/60.3%	0.46/62.6%	0.59/68.7%
Assisted	0.48/62.9%	0.54/67.8%	0.60/70.2%

Table 3. Global inter-transcriber agreement results contrasted with results reported by other studies. The numbers in the column **Pitch Accents** are the κ index and the pairwise inter-transcriber rate (as a percentage). **T** is the number of labelers, **W** is the size of the corpus in words and **S** is the number of speaking styles. *(na)* means the information is not available. The last rows of the table have been extracted from [3].

CORPUS	**T**	**W**	**S**	**Pitch Accents**
This work unassisted	3	299	1	0.51/63.9%
This work assisted	3	383	1	0.55/67.0%
Cat-ToBI [3]	10	264	4	0.462/61.17%
Am_ToBI(fe)[12]	4	644	2	0.69 / 71%
Am_ToBI(ma)[12]	4	644	2	0.67 / 72%
E_ToBI[11]	26	489	4	na / 68%
E_ToBI[16]	2	1594	1	0.51 / 86.57%
G_ToBI[6]	13	733	5	na / 71%

Table 4. Transcribers' use of the different pitch accent options expressed as a percentage, for the assisted scenario.

	First	Sec.	Third	Other	Doubt	Empty
T1	71%	20%	7%	1%	0.4%	0.0%
T2	57%	27%	3%	10%	0.4%	3.6%
T3	67%	18%	4%	11%	1.2%	2.8%

67% respectively. As for the option *Doubt*, the transcribers barely use it, which reflects self-confidence in their judgements. Finally, the label *Empty* corresponds to words with more than one stress.

Table 5 contains the inter-transcriber agreement with respect to the original labeling of the Boston Corpus. T1 has the highest agreement rate, which evidences that she is not only well-trained but also more experienced than the other two labelers.

Table 5. Consistency with the original labeling of the Boston Corpus expressed as kappa index/pairwise inter-transcriber agreement. T1, T2 and T3 represent the transcribers, and BC is the original transcriber of the Boston Corpus.

	T1-BC	T2-BC	T3-BC
Un-assisted	0.62/74.8%	0.50/63.4%	0.53/66.3%
Assisted	0.57/70.9%	0.50/63.6%	0.52/66.2%

Table 6. Consistency with the automatic labeling expressed as kappa index/pairwise inter-transcriber agreement. T1, T2 and T3 correspond to the transcribers. BC is the original transcriber of the Boston Corpus. AS is the automatic system classifier.

	BC-AS	T1-AS	T2-AS	T3-AS
Un-assisted	0.56/71.8%	0.55/71.8%	0.40/57.5%	0.44/61.9%
Assisted	0.48/66.5%	0.57/72.4%	0.41/58.4%	0.52/67.8%

The results presented in Tables 2, 3 and 5 demonstrate that computer assisted prosodic labeling introduces a bias into the labeling process of the human transcriber. Table 5 shows that the presence of automatic labels has an effect on the human experts: T1 reduces her agreement rate with respect to the original labeling. As can be observed in tables 2 and 3, both the inter-transcriber consistency and the global consistency increase because the labelers are likely to be influenced by automatic tagging.

In order to compute the savings in labeling time when the assisted scenario is used, we computed the ratio between labeling time and real time of the utterances that have been processed. The results obtained are not identical for the three human transcribers: the labeler T1, with the highest level of expertise, did not improve her labeling time, the mean value of the ratio being 55.3 if assisted by the classifier and 55.4 in the unassisted scenario. Greater time savings are observed, however, for the other labelers: the ratio is 55.95 in the assisted scenario and 67.5 when they were not computer assisted (these results show statistically significant differences when the Student's t-test is used: p-value = 0.026). These figures mean that one of these labelers could end her task about three days

before when tagging an one hour duration corpus working four hours per day (more than four hours tagging ToBI is not recommended).

Table 6 illustrates the consistency of the automatic labeling compared to the manual labelers' judgements: the value AS (automatic system) represents the first option of the three pitch accents proposed in the assisted scenario. The automatic predictions had an agreement rate as accurate as manual labelers with regard to the original tagging of the Boston Corpus (BC). In fact, only T1 has higher rates: 74.8% vs. 71.8% in the unassisted scenario and 70.9% vs. 66.5% in the assisted one. Taking into account that automatic labels can be enriched either with a degree of certainty of the prediction or with other alternative labels, we can conclude that the technique used in the automatic classifier mirrors the behaviour of the human transcriber, whose tagging, far from being utterly reliable, often results in inter-transcriber disagreement.

4 Conclusions

In this paper we have presented a computer-assisted technique for prosodic labeling of speech corpora. An automatic classifier provides the human transcriber the most likely accent types for each word, and the transcriber chooses or validates the most appropriate label.

The system has proved to be efficient, with a reduction of about 17% in manual transcription time. The effectiveness of this automatic classifier is also evident, since it reaches a success rate with respect to the original labeling which is similar to, or even higher than, those obtained by the human transcribers. Moreover, global inter-transcriber consistency improves when automatic predictions offered by the classifier are used as reference.

References

1. Ananthakrishnan, S., Narayanan, S.: Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. IEEE Transactions on Audio, Speech, and Language Processing 16(1), 216–228 (January 2008)
2. Beckman, M., Hirschberg, J., Shattuck-Hufnagel, S.: The original ToBI system and the evolution of the ToBI framework. In: Jun, S.A. (ed.) Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 9–54. Oxford University Press, New York (2005)
3. Escudero, D., Aguilar, L., Vanrell, M., Prieto, P.: Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system. Speech Communication (54), 566–582 (2012)
4. Fleiss, J.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)
5. González-Ferreras, C., Escudero-Mancebo, D., Vivaracho-Pascual, C., Cardeñoso-Payo, V.: Improving Automatic Classification of Prosodic Events by Pairwise Coupling. IEEE Transactions on Audio, Speech and Language Processing 20(7), 2045–2058 (September 2012)

6. Grice, M., Reyelt, M., Benzmüeller, R., Mayer, J., Batliner, A.: Consistency in Transcription and Labelling of German Intonation with GToBI. In: Proceedings ICSLP. pp. 1716–1719 (1996)
7. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. The annals of Statistics 26(2), 451–471 (April 1998)
8. Herman, R., McGory, J.: The conceptual similarity of intonational tones and its effects on intertranscriber reliability. Language and Speech 45, 1–36 (2002)
9. Ni, C.J., Liu, W., Xu, B.: Automatic prosodic events detection by using syllable-based acoustic, lexical and syntactic features. In: Proceedings of Interspeech 2011. pp. 2017–2020 (2011)
10. Ostendorf, M., Price, P., Shattuck, S.: The Boston University Radio News Corpus. Tech. rep., Boston University (1995)
11. Pitrelli, J.F., Beckman, M.E., Hirschberg, J.: Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Proceedings of ICSLP. pp. 123–126 (1994)
12. Syrdal, A., McGory, J.: Inter-transcriber reliability of ToBI prosodic labeling. In: International Conference on Spoken Language Processing (ICSLP). pp. 235–238 (2000)
13. Syrdal, A.K., Hirschberg, J., McGory, J., Beckman, M.: Automatic ToBI prediction and alignment to speed manual labeling of prosody. Speech Communication 33, 135–151 (2001)
14. Wightman, C., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing 2(4), 469–481 (October 1994)
15. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (December 2004)
16. Yoon, T., Chavarría, S., Cole, J., Hasegawa-Johnson, M.: Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In: Proceedings of Interspeech, Jeju. pp. 2729–2732 (2004)