Analysis of bias introduced in label assignment by computer assisted prosodic labeling

David Escudero[1], *Francisco Vizcaino*[2], *Mercedes Cabrera*[2], *Eva Estebas-Vilaplana*[3]

[1]Department of Computer Science, Universidad de Valladolid, Spain
[2]Department of Modern Philology, Universidad de Las Palmas Gran Canaria, Spain
[3]Department of Modern Languages, Universidad Nacional de Educación a Distancia, Spain
descuder@infor.uva.es, fvizcaino@dfm.ulpgc.es

Manual prosodic labeling is a costly task from a resource point of view 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 [1]. Offering the human labeler automatic predictions for them to correct or validate is a useful strategy that allows the transcriber to speed labeling. 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 the present paper we demonstrate that assisting the transcribers with ToBI prosodic labels assigned by an automatic classifier implies not only a reduction in manual transcription time but also an improvement in consistency among transcribers.

We have used an automatic classifier of ToBI prosodic events for which pitch accent classification accuracy of 70.4% has been reported [2]. One of the reasons why it is difficult to overcome this recognition rate is the high level of uncertainty aforementioned concerning the labelers' judgements. A study has been conducted which supports this fact empirically [3] in which different transcribers are asked to say what pairs of labels they find most confusing. Furthermore, some ToBI labeled corpora like [4] include notes of the transcribers stating that a second label could also be used for tagging a given accent. Additionally, in [5] we also identify the most confusing pairs of labels in inter-transcriber consistency tests. In the present work, the classifier presented in [2] has been modified to assist the manual labelers by offering them various alternative pitch accents –or absence of pitch accent– for each word. They are asked to either select the most appropriate label or, in case none of them seems to be adequate, provide one of their own. Our goal is thus for the automatic classifier to reproduce the uncertainty exhibited in the labelers' judgements. The effect of this procedure on global inter-transcriber consistency is then analysed.

The speech corpus used both for training and for testing the automatic classifier is the Boston University Radio News Corpus [4]. We have also used it to contrast the automatic prediction with the judgements of an expert labeler team. The three transcribers who participate in this study have ample experience with the ToBI labeling system. They are requested to perform the tagging task in two different scenarios, with and without automatic prosodic labeling. In the assisted scenario, the manual labeler is confronted with TextGrid files containing five tiers: one with the orthographic transcription, three with different ToBI labels, and one tier which is empty. The transcribers have previously been informed that the labels in the tiers are ranked, the one in the top tier being the most probable according to the automatic classifier. Transcribers have to fill in the bottom tier with a number indicating which of the above tiers contains the most appropriate label (see figure 1); as stated before, if none of them seems adequate, the transcribers supply their own label. In the unassisted scenario, only two tiers are provided: one with the orthographic transcription and one empty tier to be filled in by the labelers.

Table 2 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 1 shows that consistency increases in each pair, reaching more than 5 percentage points in the pair T1-T3.

Table 3 displays the use made of the different options by the transcribers. As can be seen, they select primarily 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 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.

The results presented in Tables 2, 1 and 5 demonstrate that computer assisted prosodic labeling introduces bias into the label assignment process by 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 1, both the inter-transcriber consistency and the global consistency increase because the labelers are likely to be influenced by automatic tagging.

Table 4 illustrates the consistency of the automatic labeling compared to the manual labelers' judgements: the value in column AS (automatic system) represents the first option of the three pitch accents proposed in the assisted scenario. The automatic predictions can have an agreement rate as accurate as that of the manual labelers with regard to the original tagging of the Boston Corpus (row 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.

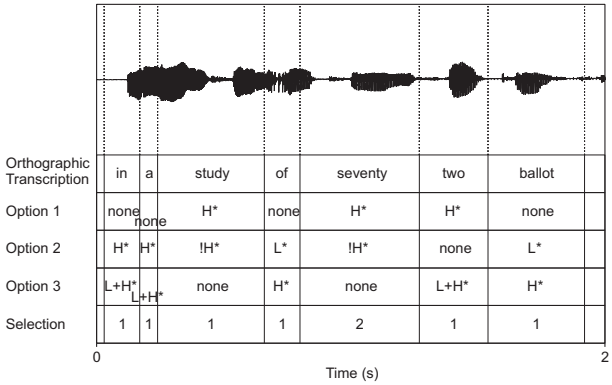

Figure 1: Praat interface in the assisted labeling scenario.

This work has been partially supported by Ministerio de Ciencia e In- novacion, Spanish Government (Glissando projects FFI2008-04982-C003-02 and FFI2011-29559-C02-01,2) and by Consejería de Educación de la Junta de Castilla y León (project VA322A11-2)

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

1. References

[1] A. K. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody," *Speech Communication*, vol. 33, pp. 135–151, 2001.

[2] C. Gonzalez, D. Escudero, C. Vivaracho, and V. Cardenoso, "Improving automatic classification of prosodic events by pairwise coupling," *IEEE Transaction on Audio, Speech and Coding*, no. 0, pp. –, in press.

[3] 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.

[4] M. Ostendorf, P. Price, and S. Shattuck, "The Boston University Radio News Corpus," Boston University, Tech. Rep., 1995.

[5] D. Escudero, L. Aguilar, M. del Mar Vanrell, and P. Prieto, "Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system," *Speech Communication*, no. 54, pp. 566–582, 2012.

[6] A. Syrdal and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling," in *Proceedings of ICSLP*, 2000, pp. 235–238.

[7] J. F. Pitrelli, M. E. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *Proceedings of ICSLP*, 1994, pp. 123–126.

[8] T. Yoon, S. Chavarría, J. Cole, and M. Hasegawa-Johnson, "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," in *Proceedings of Interspeech, Jeju*, 2004, pp. 2729–2732.

[9] M. Grice, M. Reyelt, R. Benzmüeller, J. Mayer, and A. Batliner, "Consistency in Transcription and Labelling of German Intonation with GToBI," in *Proceedings ICSLP*, 1996, pp. 1716–1719.

[10] S. Jun, S. Lee, K. Kim, and Y. Lee, "Labeler agreement in transcribing Korean intonation with K-ToBI," in *Proceedings of ICSLP*, vol. 3, 2000, pp. 211–214.

Multiclass decision

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 [5]	10	264	4	0.462/61.17%
Am_ToBI(fe)[6]	4	644	2	0.69 / 71%
Am_ToBI(ma)[6]	4	644	2	0.67 / 72%
E_ToBI[7]	26	489	4	na / 68%
E_ToBI[8]	2	1594	1	0.51 / 86.57%
G_ToBI[9]	13	733	5	na / 71%
K_ToBI[10]	21	153	5	na / 52.2%

Table 2: Global inter-transcriber agreement results contrasted with results reported for 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 labellers, **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 [5]

	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%

Table 3: Transcribers' use of the different pitch accents expressed as a percentage.

Unassisted scenario:

	BC	**T1**	**T2**	**T3**	**AS**
BC		0.62/74.8%	0.50/63.4%	0.53/66.3%	0.56/71.8%
T1			0.44/60.3%	0.46/62.6%	0.55/71.8%
T2				0.59/68.7%	0.40/57.5%
T3					0.44/61.9%
AS					

Assisted scenario:

	BC	**T1**	**T2**	**T3**	**AS**
BC		0.57/70.9%	0.50/63.6%	0.52/66.2%	0.48/66.5%
T1			0.48/62.9%	0.54/67.8%	0.57/72.4%
T2				0.60/70.2%	0.41/58.4%
T3					0.52/67.8%
AS					

Table 4: Inter-transcriber agreement 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.

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