Visualizing tool for evaluating inter-label similarity in prosodic labeling experiments

David Escudero-Mancebo[1], *Eva Estebas-Vilaplana*[2]

[1]Department of Computer Sciences, University of Valladolid, Valladolid, Spain
[2]Department of Modern Languages, Universidad Nacional de Educación a Distancia, Madrid, Spain
descuder@infor.uva.es, eestebas@flog.uned.edu

Abstract

This paper presents a technique that allows us to detect similarities among prosodic labels used to describe pitch accents within the ToBI framework. The inter-label proximity is determined empirically as a result of the evidence obtained in contingency tables of inter-transcriber agreement tests and in the confusion matrices used in automatic prosodic labeling experiments. This tool may be useful to decide which labels can be grouped together when a simplified representation is required.

Index Terms[1]: Prosodic labeling, ToBI pitch accents, tonal representation

	none	H*	!H*	H+!H*	L+H*	L+!H*	L*	L*+H	L*+!H
none	3288	150	75	20	23	2	46	1	0
H*		1032	55	19	304	50	52	44	1
!H*			248	39	11	32	24	0	5
H+!H*				48	1	4	10	0	0
L+H*					568	53	14	35	2
L+!H*						78	5	11	8
L*							218	21	18
L*+H								32	6
L*+!H									8

Table 1: Contingency matrix of an inter-transcriber consistency test (obtained from [6]).

1. Introduction

Prosodic labeling is becoming an important task in intonational studies since it triggers information that can be useful in several applications of speech technology. For example, it provides information on which parts of the message are highlighted or accented. In Automatic Speech Recognition, knowing which syllable is accented in a given word can help lexical disambiguation. In Dialog System, the identification of the focalized or highlighted items is crucial to interpret the message from a semantic or pragmatic perspective. In Text to Speech, the correspondence between prosodic form and function is fundamental to determine the expressivity of the message. Nowadays, one of the most popular systems for prosodic annotation is ToBI which distinguishes eight pitch accents for English [1]. Each pitch accent is described in terms of its linguistic function (phonological perspective) and its phonetic form which combines H and L tones.

The automatic identification of ToBI accents has been recently investigated in several studies [2][3][4][5]. The binary decision presence-absence of accent is easily evaluated and achieves over 90% identification rate [2][3]. In multiclass scenarios where different accents are assessed, the application of new techniques of classification allows to obtain a 70.8% rate [5]. It is difficult, however, to improve these results since there are certain ToBI accents that cause a high degree of uncertainty among transcribers. In [6] this observation is empirically assessed by analyzing the results of a survey which reveals the opinions of different transcribers about which tones can be more easily confused. The Boston Radio News Corpus [7] itself includes annotations from the transcribers arguing that sometimes the same accent can be associated to different labels. The existence of accent types that are rarely used as opposed to others which are assigned more frequently causes further difficulties in automatic classification, as stated in [8].

One of the proposals to overcome these problems and obtain higher identification rates in prosodic labeling has been to reduce the number of labels or pitch accent types. In [9], for instance, the ToBI labels have been simplified to four: no accent, High, Low and Downstepped. In [4], the downstepped accents have been assimilated to their non-downstepped counterparts. This paper presents a tool that may help to decide which ToBI labels can be reduced or simplified. The use of multidimensional scaling allows us to generate 2D displays that can help the experts to determine which labels can be grouped together as a result of their proximity.

The assignment of inter-label similarity will be based on empirical evidence rather than on a-priori judgments. In section 2.1 we show that our results are based on confusion matrices obtained in inter-transcriber agreement tests as well as on automatic labeling experiments. The multidimensional scaling techniques were already used for the same purpose in [10]. In this paper we present an improved version of previous investigations where the information about the relative relevance of each label is taken into account (sections 2.2 and 2.3). The graphs obtained in this study corroborate some of the a-priori judgments on inter-label similarity (section 3.1). Furthermore, they also contribute to evaluate the validity of the inter-label groupings proposed in the literature and formulate alternative proposals (section 3.2). The last experiment allows us to contrast empirically the efficiency of the proposed groupings (section 3.3).

2. Experimental procedure

2.1. The input

In [6], an experiment is described in which four transcribers label the same corpus with prosodic ToBI tags. The goal of that work was to study the inter-transcriber consistency with respect to a series of a priori judgments about the inter-symbol

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

		Automatic Classification Error	Manual Labeling Disagreement
H*	L+H*	27.26%	26.64%
H*	none	22.71%	13.15%
H*	!H*	13.17%	4.82%
!H*	none	9.75%	6.57%
L*	none	4.39%	4.03%
H*	L+!H*	3.51%	4.38%
H+!H*	none	3.22%	1.75%
!H*	L+!H*	2.88%	2.80%
H*	H+!H*	2.71%	1.67%

Table 2: Most common confusions in ToBI labeling experiments (extracted from [5]). The figures in the *Automatic Classification Error* column have been derived from the confusion matrix of a pairwise classifier trained with the data of the Boston Radio News Corpus [7]. The figures in the *Manual Labeling Disagreement* column have been obtained for the inter-transcriber test described in [6] (derived from the *all labelers-pooled* matrix in [6]).

similarity. A lateral result of the experiment is the contingency matrix displayed in Table 1 where the value of the cell in row i and column j is the number of times that one of the transcribers uses tag i for labeling an accent at which a different transcriber assigned tag j (from now onwards $n_{i,j}$). The information of this matrix will be transformed into inter-label distances for its visualization as explained in the following section.

In [5], an automatic prosodic labeling tool is presented. The tool is based on pairwise classifiers combined with expert fusion strategies. When the system is trained with the Boston Radio News Corpus [7], 70.8% of accuracy is reached in the classification of the different pitch accents types. In the automatic labeling experiment, a confusion matrix is obtained. Each cell of this confusion matrix represents the number of times the system predicts the label i when the label of the testing corpus is j. Once again, this confusion matrix will be the input of the visualizing process to display the distances between the different ToBI symbols, as will be explained in the next section.

In [5] Table 2 is included to show that the most common confusions obtained in the automatic labeling process are the same as the ones observed in the inter-transcriber consistency test reported in [6]. The visualization of the inter-label distances will show that there are also important differences between both processes.

2.2. Multidimensional scaling

Multidimensional scaling (MDS) is a set of related statistical techniques often used in information visualization for exploring similarities or dissimilarities in data [11]. Generally, the data to be analyzed is a collection of I objects on which a distance function is defined, $\delta_{i,j}$ = the distance between i^{th} and j^{th} objects.

These distances constitute the entries in the dissimilarity matrix

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix}. \quad (1)$$

such that $\delta_{i,i} = 0$, $\delta_{i,j} \geq 0$ and $\delta_{i,j} = \delta_{j,i}$. The goal of MDS is, given Δ, to find I vectors $x_1, \ldots, x_I \in \mathbb{R}^N$ such that

$$|x_i - x_j| \approx \delta_{i,j} \quad \forall i, j \in I. \quad (2)$$

Thus, MDS attempts to find a correspondence between the I objects and $\mathbb{R}^N$ such that distances are preserved. If the dimension N is chosen to be 2 or 3, we may plot the vectors x_i to obtain a visualization of the similarities between the I objects.

There are various approaches to determining the vectors x_i as they are not unique. MDS is formulated as an optimization problem to be solved numerically, where $(x_1, \ldots, x_I)$ is a minimizer of the cost function:

$$\min_{x_1,\ldots,x_I} \sum_{i<j} (|x_i - x_j| - \delta_{i,j})^2. \quad (3)$$

The obtained eigenvector and eigenvalues are used for displaying the plots [12] so that the distances in the Δ matrix are projected into the distances between I representative points. In this work, the command `cmdscale` of the software R [13] has been used. This is an implementation of the classical principal coordinates analysis for obtaining the eigenvalues from the data matrix.

2.3. Visualizing the inter-label distances

The $n_{i,j}$ values of the matrices of contingency can be interpreted as the confusion between the i and j pair of symbols. The higher $n_{i,j}$ the greater the confusion between the pair of symbols. By making $\delta_{i,j} = (max - n_{i,j} + 1)/max \; \forall i, j = 1..c$ with c indicating the number of labels and max is the maximum $n_{i,j}$ (the term $+1$ is used to avoid distances equal to 0). The Δ matrix can be obtained to be displayed by using MDS techniques. As the $n_{i,j}$ term increases, the $\delta_{i,j}$ value decreases and, as a consequence, the symbols get closer in the MDS plot.

The distances between the symbols on the MDS plot are representative of the confusion between them. Two symbols appear close to each other when different labelers have frequently assigned these symbols to the same event in the transcription procedure or when the automatic labeling system predicts the symbol for items previously labeled with the other symbol in the testing corpus. MDS techniques allow a set of x_i vectors with $i = 1..I$ to be obtained such that each x_i represents a class of symbols. The distance between the vectors is assumed to be proportional to the confusion between the symbols. Once again, we use a 2D plot to display the distances between the ToBI symbols.

Apart from the distances between the ToBI labels, the contingency matrices bring additional information that is also interesting to take into account. On the one hand, the total amount of i labels that are used or predicted in the experiments, computed as $n_i = \sum_{j=1}^{c} n_{i,j}$ measures the total use of the i symbol as an indicator of its *relevance*. On the other hand, the relative weight of $n_{i,i}$ with respect to n_i is representative of the *consistency* of the i label.

In order to visualize the *consistency* and the *relevance* of the ToBI labels in the MDS plot, we place the tag in an adaptable graphic symbol. The radius of the symbol is proportional to the *relevance* of the label and the color of the symbol represents its *consistency*.

The *relevance* index will be $r_i \propto \sum_{j=1}^{c}(n_{i,j} + n_{j,i}) - 2 * n_{i,i}$. We use in this work a linear function with maximum and minimum thresholds for limiting the size of the circles.

The *consistency* index will be $g_i = 1 - n_{i,i}/n_i$ so that the closer the gray scale of the circle to 1 (white color) the more

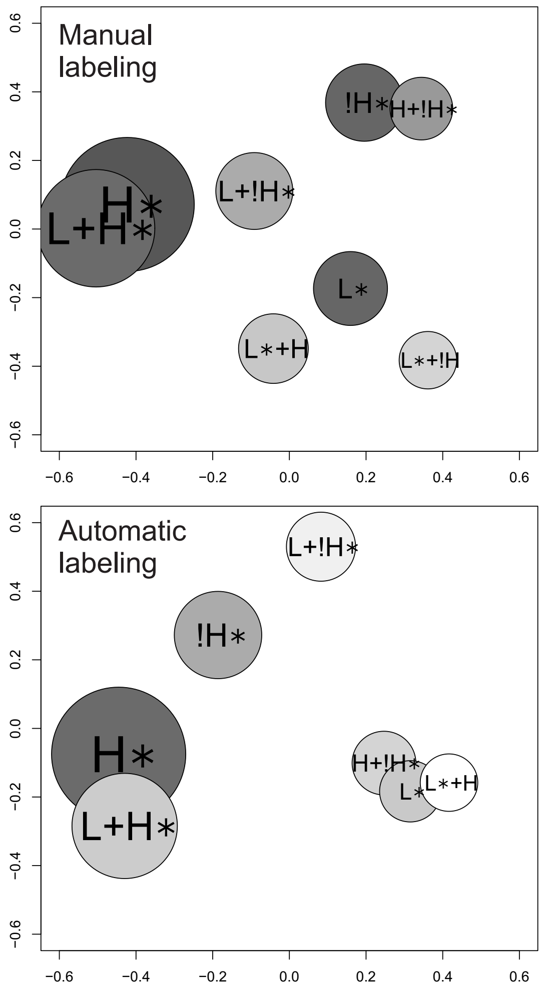

Figure 1: MDS diagrams of the inter-label confusion.

the label is confused. Similarly a darker circle implies that $n_{i.i}$ is close to n_i meaning that the confusion of this symbol is relatively low. When $n_i = n_{i,i}$ the color of the symbol is black indicating that the symbol is never confused.

3. Results

3.1. Visualizing the inter-label confusion

Figure 1 shows the multidimensional scaling plots that project the inter-label distances. The upper plot represents the distances obtained in an inter-transcriber consistency test. The lower plot represents the distances derived from the confusion matrix of the automatic labeling experiment.

Both plots exhibit some similarities. H* is the biggest and darkest circle. This means that H* is the most frequently used symbol and, in relative terms, the less confused one both by the manual ToBI transcribers and by the automatic labeling system. The pair L+H* and H* is represented by overlapped circles, indicating that this pair is highly confused. The labels containing a low tone (L*, L*+H and L*+!H) tend to be grouped naturally, meaning that they are interchanged many times.

Apart from these inter-plot similarities, there are also clear differences between the plots. The behavior of the downstepped symbols seems to be different. In the manual inter-transcriber experiment, !H* and H+!H* tend to be confused frequently, but the automatic system fails in that respect and places the symbol H+!H* close to the L* symbols. The label L+!H* is close to !H* in both cases. However, the manual labelers place it close to its no downstepped counterpart L+H* as well.

The shading of the circles reflect another clear difference between the manual labeling inconsistencies and the automatic labeling errors. The automatic labeler unbalances more the predictions of the symbols. Thus, the difference of the intensity of the symbol H* with respect to the rest of symbols is higher in the automatic labeling plot. Indeed, the symbols L+!H* and L*+H are very rarely used. The explanation for this fact is that the classifier tends to get specialized in the most populated class in order to increase its global performance. The use of Multidimensional Scaling plots evidences this behavior which triggers the inevitable manual revision of the automatic predictions to validate them.

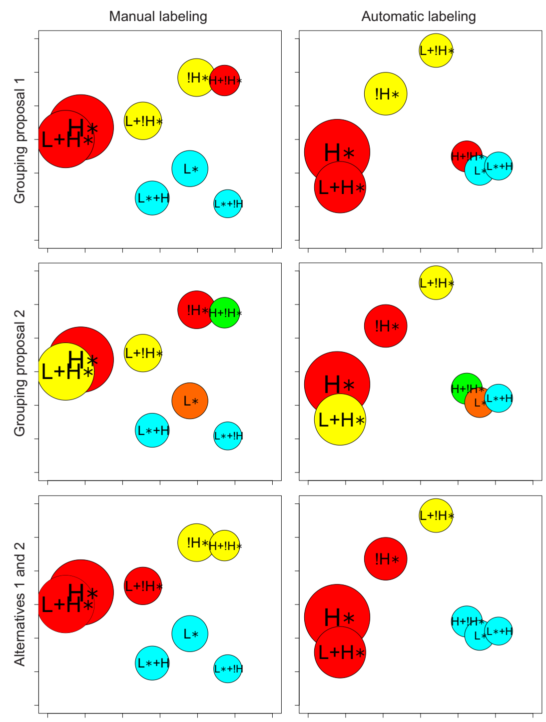

Figure 2: MDS diagrams representing the classes that are merged together in different proposals and alternatives. Same color means the classes that are collapsed into a common one.

3.2. Grouping proposals and alternatives

In [9] authors propose to group the ToBI tone labels into four categories: No accent, $High \equiv \{L+H*, \quad H*, \quad H+!H*\}$, $Downstepped \equiv \{L+!H*, \quad !H*\}$ and $Low \equiv \{L*, \quad L*+H, \quad L*+!H\}$. This classification was also used by [14] and more recently by [15]. The upper graphic in Figure 2 shows that the classification is coherent with the inter-transcriber agreement rates. However the symbol $H+!H*$ can be problematic.

An alternative simplification where the downstepped symbols are collapsed with their normal counterpart is presented in [4]. That is $\{L+H*\}' \equiv \{L+H*, L+!H*\}$,$\{H*\}' \equiv \{H*, !H*\}$, $\{L*+H\}' \equiv \{L*+H, L*+!H\}$. This strategy seems to be more risky according to the plots displayed in the

Grouping Proposal	WONAC			WNAC		
	N	Ac	FM	N	Ac	FM
No Grouping	7	58.4	57.3	8	45.0	42.5
Grouping 1	3	74.9	68.1	4	71.1	69.3
Grouping 2	5	67.1	60.4	6	66.4	64.7
Alternative 1	3	75.7	69.8	4	71.6	69.8
Alternative 2	3	86.9	82.3	4	76.6	75.1

Table 3: Classification results in terms of the accent grouping proposal. $WONAC$ means *Without no accent class*. $WNAC$ means *With no accent class*. N is the number of classes. Ac is the accuracy of the classifier. FM is the f-measure of the classifier. *Grouping 1* is the one described in [9]. *Grouping 2* is the one proposed by [4]. *Alternative 1* is the one arising from the inter-transcriber confusion tests. *Alternative 2* is obtained with the results of the automatic classifier.

middle row of Figure 2. This classification is more expressive than the one presented previously as it has five different types of accents instead of only three.

The third row of Figure 2 presents two new proposals that are guided by the results obtained in the inter-transcriber tests and in the automatic classification experiments respectively. The classification has been done according to the inter-label distance observed in the diagrams. *Alternative 1* collapses three classes so that $Class1 \equiv \{L*, L*+!H, L*+H, \}$, $Class2 \equiv \{H*, L+!H*, L+H*\}$, and $Class3 \equiv \{!H*, H+!H*\}$. *Alternative 2* collapses two classes so that $Class1 \equiv \{L*, L*+H, H+!H*\}$ and $Class2 \equiv \{H*, L+H*, !H*\}$.

3.3. Efficiency of the grouping proposals

Table 3 shows the classification accuracy after applying the grouping proposal in the original data. We use the implementation of J48 decision tree of the Weka tools in [13]. The input features are the ones described in [2] including F0, energy, duration and POS Tags. Contrary to [5], we do not use Tilt or Bézier features, nor fusion of experts or identification of sequences. The results could improve in all cases if we changed the classifier but this is not the aim of this work. The goal of this experiment was not to test the efficiency of the classifier but to contrast the efficiency of the different merging proposals.

Any of the grouping strategies considerably improves the labeling accuracy with respect to the one obtained when the original classification is used (compare the No Grouping row in Table 3 with the other rows). *Grouping 2* is difficult to compare with the rest because it uses more classes. *Alternative 1* and *Alternative 2* both improve *Grouping 1* (75.7% and 86.9% vs. 74.9% respectively). The ranking remains when the class *no Accent* is entered (columns WONAC vs. WNAC in Table 3).

4. Conclusions

In this paper we have presented a procedure that allows to evaluate the proximity and similarity of ToBI labels. The findings obtained in this paper, however, can also be applied to other systems of intonational annotation. Multidimensional Scaling graphs have been used to show the relevance of the different pitch accent types. Based on these plots, experts can justify inter-label groupings that can be used in simplified representations.

The process has been developed using contingency tables of inter-transcriber agreement tests and the confusion matrices obtained in an automatic prosodic labeling experiment. Both procedures present grouping alternatives that are proved to be more efficient than former groupings proposed in other works.

This paper does not examine alternative classifications. Its main contribution is to support transcribers in deciding how to simplify prosodic labeling systems. The alternative classifications we obtained seem to be operative, however, we should bear in mind that they depend on the original data and hence may vary in other experiments. Even though other confusion matrices can trigger different proposals, the visualization and decision processes remain the same.

Finally, the application of this technique is not restricted to ToBI labels, but could be applied to other types of prosody transcriptions, and probably to many types of annotation; to be shown in future work.

5. References

[1] M. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford University Press, New York, 2005, pp. 9–54.

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

[3] C.-J. Ni, W. Liu, and B. Xu, "Automatic prosodic events detection by using syllable-based acoustic, lexical and syntactic features," in *Proceedings of Interspeech 2011*, 2011, pp. 2017–2020.

[4] A. Rosenberg, "Classification of Prosodic Events using Quantized Contour Modeling," in *HLT-NAACL*, 2010, pp. 721–724.

[5] C. González, D. Escudero, C. Vivaracho, and V. Cardeñoso, "Improving automatic classification of prosodic events by pairwise coupling," *IEEE Transaction on Audio, Speech and Language Processing*, p. in press, 2012.

[6] 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.

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

[8] C. G. Ferreras, C. Vivaracho-Pascual, D. E. Mancebo, and V. C. noso Payo, "On the automatic tobi accent type identification from data," in *INTERSPEECH 2010*, 2010, pp. 142–145.

[9] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, 1996.

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

[11] J. Kruskal and M. Wish, *Multidimensonal Scaling*. Sage University Paper series on Quantitative Application in the Social Sciences, Beverly Hills, 1978.

[12] I. Borg and P. Groenen, *Modern Multimensional Scaling: theory and applications*. Springer-Verlag New York, 2005.

[13] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[14] X. Sun, "Pitch accent prediction using ensemble machine learning," in *International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 16–20.

[15] G. Levow, "Context in Multi-lingual Tone and Pitch Accent Recognition," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2005, pp. 1809–1812.