Analysis of Cat-ToBI indices intertranscriber inconsistencies: implications for automatic labelling

David Escudero-Mancebo, Lourdes Aguilar

Department of Computer Science, Universidad de Valladolid, Spain
Department of Hispanic Studies, Universidad Autonoma de Barcelona, Spain
descuder@infor.uva.es, Lourdes.Aguilar@uab.cat

Abstract

In this paper we present an experience to measure the inter-transcribers consisteny where a number of observer have been required to identify ToBI events in the same set of sentences. We computed the pairwise transcribers agreement with its corresponding confusion matrix and the kappa coefficients. The goal was to identify the main sources of confusion resulting: (1) bad trained observers (2) problematic symbols. The identification of those problematic symbols supports the practical decision to merge them into an alternative class when automatic approaches to ToBI labelling are focused; in this case for ToBI break indices.

Index Terms: prosody, ToBI, inter-transcriber consistency, automatic recognition

1. Introduction

ToBI is a standard for representing and labelling prosodic events including tones (accent tones and boundary tones) and breaks [1]. 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 number 0 represents no boundary, 4 represents a full intonation phrase boundary and the rest of indices are breaks with intermediate strength. In this paper we focus on breaks as the genesis of this work was the potential interest of breaks for the representation of the utterance rhythmic structure with applications in text-to-speech systems.

ToBI has been implemented for several languages including English, German and Japanish. Concerning to Iberian languages, it exists active groups responsible for the Cat-ToBI and Sp-ToBI for Catalan and Spanish respectively. The need of a reference corpus similar as the ones existing for other languages (e.g. the Boston Radio Corpus for English [2]) 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 containing Radio news recordings and spontaneous dialogs.

Labelling a corpus with ToBI tags is an expensive procedure. In [3] 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. [4] or [5] are good examples of the state of art on automatic labelling of ToBI events. For Catalan we presented a work for labelling break indices [6]. In that work we reduced 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 [4] or [5] that combine the different type of accent tones transforming the labelling problem into a binary one to decide weather an accent is present or not.

In this work we show that grouping different labels is a coherent procedure according to the diversity of judges observed in an inter-transcriber experiment. We present an experiment where different labellers are required to assign different ToBI tags in the same reduced set of sentences. Results seems to indicate that some of the ToBI tags are easier to confuse for the labellers. The more the confusion between a pair of classes the more the evidence that this pair of classes is a good candidate to be merged. We present a tool to compute and visualize the inter-transcriber inconsistency and we discuss about the inter ToBI labels confusion values.

First we present the experimental procedure with the corpus used, next the experimental procedure indicating which metrics have been applied and the procedure to visualize information. Finally we conclude with discussion and future work.

2. Experimental Procedure

A test of labeling consistency was conducted to measure inter-transcriber consistency in the Cat-ToBI prosodic transcription system in order to asess the system and to detect if there are labels frequently confused. Twenty utterances were excerpted from four different speech styles produced by twelve different speakers and transcribed by ten labelers differing in their levels of experience with Cat-ToBI.

2.1. Speech database

To assess the labeling conventions of Cat-ToBI and to demonstrate that these conventions are applicable to various types of speech, we selected twenty utterances representing four different discourse types: spontaneous speech excerpted from the database of the Atles interactiu de l'entonacio catala[2], in particular, from the intonation survey and the Map Task dialogue corpus; radio news and text reading (from the Festcat database[7]). Twelve speakers (5 male and 7 female) produced the sentences. These sentences contained a total of 264 words, and lasted a total of 89.8 seconds. Nine of the sentences are interrogative questions, four are emphatic declarative, and the rest, neutral declarative.

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

[2]http://prosodia.upf.edu/atlesentonacio/metodologia/index-english.html

2.2. Subjects

The subjects span a variety of levels of experience with prosody and experience with Cat-ToBI ranging from absolute beginners to contributors to its development. The labelers were divided into three groups: Group 1 (Experts), Group 2 (Familiar with prosodic annotation systems), and Group 3 (Beginners, completely new to any model of intonation or prosodic transcription). Each group included four labelers, except for Group 2 which had two labelers. All subjects are native speakers of Catalan, with two dialects represented (Central Catalan, Balearic Catalan).

Each transcriber was provided with a document describing the Cat-ToBI system[8] and with the Cat-ToBI training materials[3]. The training materials contain a tutorial explaining each of the labels in Cat-ToBI, along with recorded examples of transcribed utterances. There are also exercises to practice the labels described in the text. The training materials were designed to be self-explicative. Moreover, absolute beginners attended a course (three sessions of three hours each) on the basics of AM model and the ToBI labelling systems.

All the labellers were given a document with basic instructions and a package with the sound files and the textgrids, with the Praat tool[4]. The selected speech has not been previously labeled by any of the transcribers; each transcriber worked alone on the samples and they were not allowed to discuss utterances in the experimental data-set. After they completed the transcription, their textgrid files were collected and statistics for labeler agreement were applied to the data.

2.3. Transcription procedure

The manual annotation was performed using the Praat tool. The transcribers were looking at a computer screen with a display of the signal (F0 curve and waveform) and they rely on auditive and visual information to take their prosodic decisions. The key elements to be labeled are prominence, prosodic boundary strength and pitch accent and boundary tone types.

In the ToBI framework, the transcribers have to perform the following tasks:

1. Mark syllables which are carrying a clear prominence, that is, decide if there is a pitch accent
2. If there is a pitch accent, decide the pitch accent type
3. Mark important between-word interruptions of the normal speech stream as either weak (signalling intermediate phrases) or strong breaks (that is, intonational phrases)
4. Decide the boundary tone type

2.4. Reliabilty measurements

2.4.1. Pairwise transcriber agreement

Agreement was measured by counting the number of labeling agreement 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), and the criterion is conservative: if 3 of 4 transcribers agree, only 3 of 6 pairs will match, making the agreement rate 50% (agreement = agree / (disagree + agree)).

For example, if a particular word boundary was labeled by the first transcriber as 2, by the second transcriber as 3, and by transcribers 3 and 4, as 2, the number of transcriber pairs who agree with each other is three (T1T3, T1T4, T3T4) and the number of transcriber pairs who disagree with each other is also three (T1T2, T2T3, T2T4).

We are representing this information as a confusion matrix where rows and columns index the ToBI symbol. The main diagonal is indicates the coincidences and the rest of the elements are the discrepances.

2.4.2. Kappa coefficient

Cohen's kappa[9], which works for two raters, and Fleiss' kappa[10], 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 through chance.

Agreement can be thought of as follows, 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,

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

Where $\bar{P}$ is an array measuring the agreement for the different symbols and $\bar{P}_e$ is the hypothetical probability of chance agreement. 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. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.

The inter-transcriber consistency for prominence, break strength and pitch accent and boundary tone inventory was quantified by means of kappa coefficient. According to [11], a kappa between 0.61 and 0.80 is considered to point at a substantial consistency. [12] considered a good level of agreement when the value obtained from the kappa statistic is greater than 0,7.

2.5. Visualizing the inconsistency

We use the Kappa Fleiss coefficient to obtain a symmetric matrix of distances between the ToBI events. This matrix is indexed in term of the type of break and represents the inconsistency between every pair of breaks. The more the distance the easier it was to distinguish this pair of breaks by the labellers. Given a pair of break indices, the whole set of decisions is binarized setting to `un-available data` the decisions that do not concern the selected pair and the Kappa Fleiss coefficient is computed. The more coincidences between the labellers referring to the given pair of breaks, the higher the corresponding κ value in the distance matrix.

Multidimensional scaling (MDS) is a set of related statistical techniques often used in information visualization for exploring similarities or dissimilarities in data. An MDS algorithm starts with a matrix of itemitem similarities, then assigns a location to each item in N-dimensional space, where N is specified a priori. For sufficiently small N, the resulting locations may be displayed in a graph or 3D visualisation. Multidimensional scaling will be used to display our distance matrix of break indices in a 2D plot. The closer the breaks, the more the confusion.

A similar procedure will be applied to obtain a distance matrix between the labellers. The more the agreement between a

[3] http://prosodia.uab.cat/cat_tobi/en/index.php

[4] http://www.praat.org

	B0	B1	B2	B3	B4
B0	414	164	6	23	1
B1		332	29	61	1
B2			8	22	4
B3				105	39
B4					163

Figure 1: Confusion matrices where cells compute the number of ocurrencies of the pair indexed by row and column

pair of labellers the closer will be displayed in a 2D plot. The distance between every pair of labellers will be computed as $1 - \kappa$, been κ the inter-transcriber agreement of the pair of labellers.

We use classical multidimensional scaling [13], in particular its implementation in R[5] `cmdscale` procedure. We used the Interrater Reliability and Agreement. (irr) R package to compute the kappa coeficients.

3. Results

The global intertranscriber rate of agreement is 74.49 % which is a moderate result when compared with the test performed with consolidate ToBI systems in the rates trainning process: Previous works on intertranscriber reliability of ToBI-framework systems have certified between 81% and 92% of agreement in determining pitch accents for English [12], overall mean scores of 88.9% of agreement for German [14], and agreement percentages of between 59% and 91% (depending on accent categories) for Korean ([15]).

When the intertranscriber rate of agreement is split in the corresponding confusion matrix (table 1) we see clearly that there are important differences among the indexes. Thus, breaks 0 and 4 are identified easily, meanwhile the break 2 is identified with about 10% agreement. With respect to the breaks 1 and 3, the confusion with the other indexes is also high.

To display these discrepancies, we use the kappa coefficient. Table 2(top) shows a table of the kappa Fleiss coefficient for every pair of break indices according to the procedure explained in section2.4.2. Figure 2(down) interprets the kappa Fleiss coefficients as distances to apply multidimensional scaling. We can observe that it seems to be three groups of breaks: break 0, break 4 and a third group formed by the breaks 1, 2 and 3.

We obtain a kappa coefficient of 0.666 that corresponds to a substantial agreement in the commonly used kappa scale. As it was explained in section 2.2, there are three groups of raters: Experts, Beginners and Intermediates. If we separate the rates assigned by these two groups we obtain a kappa coefficient of 0.75 in the expert group. To display these discrepancies among the taggers we refer again to section 2.4.2 to build the table and figure in 3. We remark here the important differences among the different labellers. This type of figures could potentially be used to check the labeller reliability, under the supposition, that the closer the labeller is to an expert, the more accurate his or her rates are. Thus, the labeller **i1** behaves as a goat (in biometric terminology) meanwhile others behave as sheeps, always close to an expert.

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

	B0	B1	B2	B3	B4
B0		0.715	0.667	0.639	0.746
B1			0.586	0.602	0.659
B2				0.524	0.727
B3					0.746
B4					

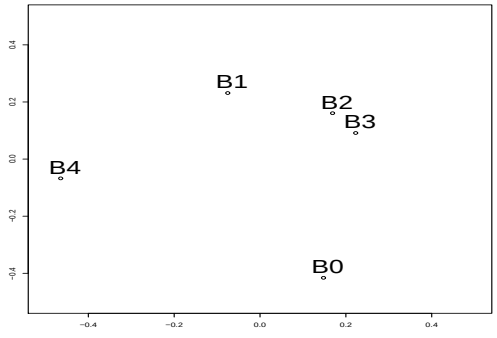

Figure 2: Multidimensional 2D plot of the the distances between de different Breaks (Kappa coefficients in the table above). B0..B4 are the different break indexes.

Figure 4 illustrates the inter transcribers confusion when the different breaks are isolated. Once again we see that the breaks 0 and 4 are the less problematic. Break 2 is very confusing. Table 1 shows that the break 2 appears very few and when it does, it is marked by 1 or 2 labellers most of the times. This observation makes it congruent the merging of break 2 with other classes in practical situations.

With respect to the break 3 and break 1 distinction, figure 4 seems to indicate that there are two groups of taggers (as marked on the figure). This cluster of labellers could be indicating the use of different criteria and it is something to analyze in future works. In the time this divergent criteria is solved, it seems congruent to merge the labels into only one reducing the number of break tags from five to only three as it was done in our previous work [6] on automatic identification of break indices.

	1	2	3	4	5	6	7	8
B0	25	16	3	5	4	5	9	48
B1	11	11	9	7	5	27	13	20
B2	24	3	1	2	1			
B3	10	6	6	11	5	4	5	4
B4	2		6	3	1			22

Table 1: Frecuency of different ToBI labels: the cell quantity is the number of times that the break index of the row was labelled by the number of labellers indicated by the column.

4. Conclusions

In this paper we have run a test of inter-transcribers consistency consisting on the ToBI labelling of a set of Catalan sentences by a number of observers.

Results show that one of the main sources of confusion has its origin in bad trained labellers which observations separates clearly from the experts labellers. Another source of noise is

	b1	b2	b3	E1	E2	i1	E3	E4
b1		0.665	0.645	0.653	0.79	0.597	0.668	0.665
b2			0.699	0.832	0.69	0.537	0.631	1
b3				0.658	0.705	0.504	0.57	0.699
E1					0.728	0.516	0.619	0.832
E2						0.547	0.727	0.69
i1							0.556	0.537
E3								0.631
E4								

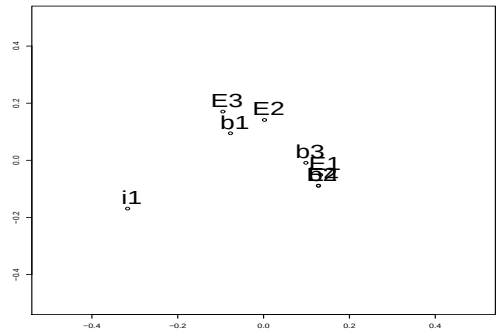

Figure 3: Intertranscriber discrepancy for Breaks among the different labellers. The Kappa indices of the table are projected in a 2D multidimensial scaling plot. E is expert, b is beginner and i is intermediate.

the use of symbols (like break index 2) with an apparently fuzzy definition leading to a scarce use in only rare situations.

The use of the visualization tools has shown to be useful to identify potential diverse tagging criteria. The 2D plots have shown to be useful to detect clusters of labellers as an evidence of possible different labelling criteria.

All these observations lead us to conclude that the merging of labels is a congruent procedure with practical advantages supported by the observation run on perceptual tests. The visualization tools permits to identify the closest symbols to be merged.

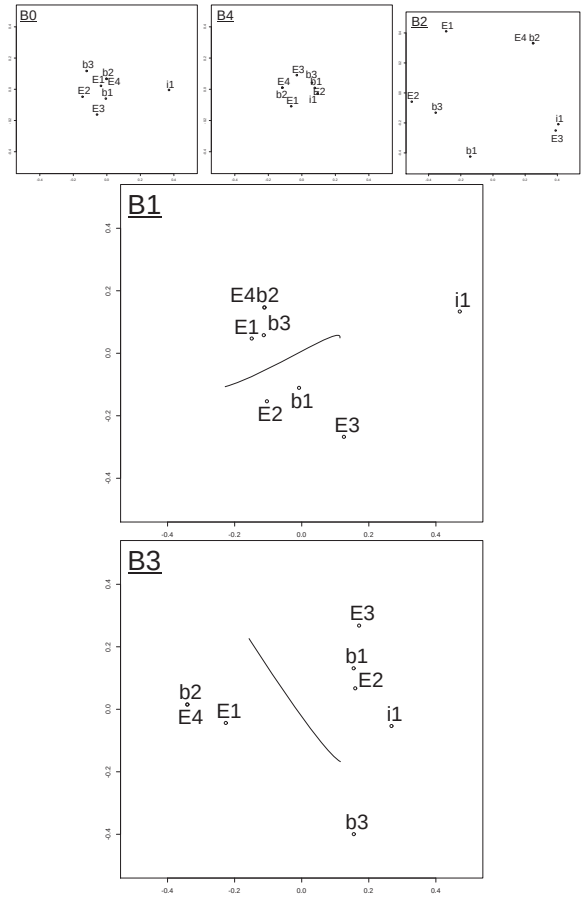

Figure 4: Intertranscriber discrepancy among the different labellers isolating the decisions for every type of Break

5. References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Perez, "Corpus and Voices for Catalan Speech Synthesis," in *Proceedings of LREC 2008*, 2008.

[8] P. Prieto, L. Aguilar, I. Mascar, F. Torres, and M. Vanrell, "L'etiquetatge prosdic Cat-ToBI," *Estudios de Fonetica Experimental*, vol. XVIII, pp. 287–309, 2009.

[9] J. Cohen, "A coefficient for agreement for nominal scales," *Education and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[10] J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[11] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[12] T. Yoon, S. Chavarria, J. Cole, and M. Hasegawa, "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," in *Proceedings of Interspeech 2004*, 2004.

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

[14] M. Grice, R. Reyelt, R. Benzsuller, and A. Batliner, "Consistency in transcription and labelling of German intonation with GToBI," in *Proceedings of ICSLP 1996*, 1996, pp. 1716–1719.

[15] S. Jun, S. Lee, K. Kim, and Y. Lee, "Labeler agreement in transcribing Korean intonation with K-ToBI," in *Proceedings of ICSLP 2000*, 2000.