On the use of a fuzzy classifier to speed up the Sp_ToBI labeling of the Glissando Spanish corpus

David Escudero-Mancebo[1], **Lourdes Aguilar-Cuevas**[2], **César González-Ferreras**[1], **Yurena Gutiérrez-González**[2] **and Valentín Cardeñoso-Payo**[1]

[1]Department of Computer Science, University of Valladolid, Spain
[2]Department of Spanish Philology, Universitat Autònoma de Barcelona, Spain
{descuder,cesargf,valen}@infor.uva.es, lourdes.aguilar@uab.es

Abstract

In this paper, we present the application of a novel automatic prosodic labeling methodology for speeding up the manual labeling of the Glissando corpus (Spanish read news items). The methodology is based on the use of soft classification techniques. The output of the automatic system consists on a set of label candidates per word. The number of predicted candidates depends on the degree of certainty assigned by the classifier to each of the predictions. The manual transcriber checks the sets of predictions to select the correct one. We describe the fundamentals of the fuzzy classification tool and its training with a corpus labeled with Sp_TOBI labels. Results show a clear coherence between the most confused labels in the output of the automatic classifier and the most confused labels detected in inter-transcriber consistency tests. More importantly, in a preliminary test, the real time ratio of the labeling process was 1:66 when the template of predictions is used and 1:80 when it is not.

Keywords: Prosodic labeling, fuzzy classifier, Sp_ToBI

1. Introduction

Prosody is a component of speech that can be useful in several applications of speech technology, as it provides information, among other things, on which parts of a message are highlighted or accented. In Automatic Speech Recognition, knowing whether syllable within a given word is accented or not can help resolve lexical disambiguation. In Dialog Systems, the identification of focalized or highlighted items can be crucial in interpreting 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.

The study of these aspects, then, requires the availability of recorded corpora with prosodic labels that identify the acoustic segments where significant prosodic events occur. The standard labeling process of corpora is costly in terms of time and resources because it requires the intervention of human coders. An additional problem is that the human factor also includes a high level of uncertainty which can cause inconsistency across labelers (Wightman, 2002). In order to confront these situations, we recently presented a methodology that makes use of fuzzy sets (Escudero-Mancebo et al., 2014). We will step forward by applying this methodology to propose an automatic classifier which generates ToBI labels in Spanish corpus assuming certain degrees of uncertainty.

Spanish prosody has been the focus of research since the seminal studies of (Navarro-Tomás, 1944). The study of Spanish intonation has recently widen thanks to several works carried out with the Autosegmental Metrical model (AM) and the follow-up Sp_ToBI system. The inventory of pitch accents and tone boundaries in the intonation of Spanish is well described (Beckman, 2002; Hualde, 2003; Sosa, 2003; Face and Prieto, 2007), and despite the distinct trends occurring in various dialects of Spanish, (Prieto and Roseano, 2010) developed a common framework of analysis of the tonal patterns and prosodic structure found in Spanish and a common set of ToBI labels to account for them.

The Glissando corpus provides an amount of information for prosodic studies similar to the one used in other languages, such as the Boston University Radio News Corpus (Ostendorf et al., 1995), dialogues such as the Buckeye Corpus (Pitt et al., 2005) and spontaneous speech such as the Corpus of Spontaneous Japanese (Maekawa, 2004). Thus, it offers a great number of contextualized sentences which make the inspection of some problems easier and, as a consequence, supports improvement proposals. Currently, the Glissando project already includes a part of the news corpus manually labelled within the Sp_ToBI framework.

The aim of this study is, in the first place, to explore to what extent the commonly accepted conventions of Sp_ToBI system are suitable for labelling a large-size corpus, such as the Glissando corpus (Garrido et al., 2013), and then propose a tool that helps to speed up the process of prosodic labelling by automatically proposing labels that take into account degrees of uncertainty.

Labeling a corpus with ToBI tags is an expensive procedure. In (Syrdal et al., 2001) it is estimated that ToBI labeling commonly takes from 100-200 times real time. To speed up the process, automatic or semiautomatic methods seem to be a suitable aid. (Ananthakrishnan and Narayanan, 2008b; Rangarajan Sridhar et al., 2008) are good examples of the state of art on non-fuzzy automatic labeling of ToBI events. A second goal of this work is, then, to present a new methodology that automatically proposes ToBI labels from the acoustic detection of prosodic events, and the application of this procedure to include Sp_ToBI labels in the Glissando corpus automatically in order to obtain a high throughput in the production of labels.This methodology is based on fuzzy sets and it preserves the complexity of prosodic phenomena and the relevance of the

CORPUS	L	W	S	Pitch Accents	Boundary Tones	Breaks
Sp_ToBI (this work)	4	108	2	0.68/78.35%	0.70/85.05%	0.76/88.63%
Cat_ToBI(Escudero et al., 2012)	10	264	4	0.462/61.17%	0.69/86.10%	0.68/77.14%
Am_ToBI(fe)(Syrdal and McGory, 2000)	4	644	2	0.69 / 71%	0.84 / 86%	0.65 / 74%
Am_ToBI(ma)(Syrdal and McGory, 2000)	4	644	2	0.67 / 72%	0.76 / 82%	0.62 / 74%
E_ToBI(Pitrelli et al., 1994)	26	489	4	na / 68%	na / 85%	na / 67%
E_ToBI(Yoon et al., 2004)	2	1594	1	0.51 / 86.57%	0.79/ 89.33%	na / na

Table 1: Global inter-transcriber agreement results for Sp_ToBI contrasted with results reported for other ToBI systems. Columns labelled *Pitch Accents*, *Boundary Tones* and *Breaks* separate results according to the respective ToBI events that have been considered. The figure in the cells are the κ index and the pairwise inter-transcriber rate (as a percentage). **L** is the number of labelers, **W** is the size of the corpus in words and **S** is the number of styles. *(fe)* is female, *(ma)* is male and *(na)* means the information is not available.

listener's perception in the intonational phonology framework, since it offers alternative labels enriched with associated degrees of uncertainty. Ultimately, it is the human transcriber who decides which of the proposed labels corresponds to the perceived tone. Section 2 presents the preparation of the Sp_ToBI labeled training corpus and section 3 describes the fuzzy classifier and results.

2. The training corpus

2.1. Sp_ToBI manual annotations

The Glissando news subcorpus contains recordings of eight different Spanish speakers each of them reading more than 36 news item (Garrido et al., 2013). For our purposes, two of these speakers were chosen, taking into account difference in gender (i.e. male and female) and reading style (i.e radio speaker and advertisement actor). The labeled corpus consists of 1100s of news reading speech recorded by two professional speakers: 12 news read by a radio professional (woman's voice) and 12 news read by an advertising professional (male voice).

The news data-set has been annotated using the Sp_ToBI labels proposed in (Estebas Vilaplana and Prieto, 2009; Estebas Vilaplana and Prieto, 2010), with the modifications advanced in (Elordieta, 2011). A phonologically-oriented prosodic annotation, such as the ToBI model, requires a wide consensus on particular aspects of a restricted speech style, such as the reading news by professionals. In this study, various methods of validating the consistency and stability of the labels assigned to the corpus were conducted: (i) periodical meetings to define a proposal that applies the Sp_ToBI to news reading; (ii) discussion and resolution of differences in transcription throughout a six-month period and (iii) validation of consistency among transcribers with an interreliability experiment.

The results of the intertranscriber consistency test can be seen in the table 1. Values of the kappa index between 0.6 and 0.8 like the ones we obtained are commonly considered as substantial agreement. These consistency rates are comparable with the ones reported in similar studies for the prosodic labeling of other corpora in different languages (see table 1). Uncertainty exists, which is the main argument that supports the use of a fuzzy classifier.

After this, one of the transcribers participating in the consistency test was recruited, and the annotation decisions were reviewed by another expert. The annotation was not considered definitive until the transcriber and the reviewer reached a consensus on the labeling. The procedure was perceptually based. The transcriber was encouraged to focus preferentially on perception. Her task consisted in listening carefully to the utterance in order to (a) mark the subjective sense of disjuncture between each pair of words and before each pause (break tier) and (b) mark prominences and tonal events (tone tier).

The labeler concluded the transcriptions of 24 news items that include a total of 3202 words. She marked a total of 2058 pitch accents and 1115 boundary tones and 1029 breaks (see Table 2 for details).

Some categories show a very low number of instances, so we decided to group them with similar types, thereby creating particular classes. To do that, we display the inter-label distance into a *Multidimesional Scaling* (MDS) 2D plot following the perspective adopted in (Escudero-Mancebo and Estebas-Vilaplana, 2012). This MDS map is built with the confusion matrix of a decision tree classifier: the more the inter-class confusion the closer the labels in the map. This plot allows experts to make a decision regarding the different categories. The closest categories on the map are good candidates to be merged into new groups. It was also assumed that there were similarities among Sp_ToBI symbols, including parentheses and their corresponding counterparts outside of the brackets (e.g. the samples (L+)H* are gathered with L+H* class). (More details in (Escudero et al., 2014)). Table 2 shows the number of samples per label. Classification rates greatly improved after performing this clustering.

In order to train the classifiers, we have applied a multifold approach that divides the corpus into two sections: 90% training and 10% test.

2.2. Acoustic prosodic features

The acoustic prosodic features that have been use to train and test the automatic labeler are similar to those reported in other experiments (Ananthakrishnan and Narayanan, 2008b):

- Frequency features: within-word F0 range, difference between maximum and average within-word F0, difference between average and minimum within-word

Pitch Accent	!H*	H*	L*	L+>!H*	L+>¡H*	L+>H*	L+!H*	L+¡H*	L+H*	none
#	103	187	94	14	8	254	275	125	998	1144

Boundary Tone	=%	!H%	¡H%	%H	%!H	H%	L%	LH%	none	
#	121	251	1	41	45	337	251	68	2087	

Breaks	BI 3	BI 4	other							
#	811	218	2173							

Table 2: Number of prosodic labels in the training corpus

F0, difference between within-word F0 average and utterance average F0.

- Energy features: within-word energy range, difference between maximum and average within-word energy, difference between average and minimum within-word energy.
- Maximum normalized vowel nucleus duration from all of the vowels of the word. Normalization is performed for each vowel type.
- Part of speech (POS): we used the POS tags that were automatically obtained using FreeLing, an open source language analysis tool suite (Padró and Stanilovsky, 2012). The EAGLES tag set for Spanish was used.

In order to model the temporal evolution of the pitch contour, we have also used Tilt (Taylor, 2000) and Bézier (Escudero et al., 2002) features, as described in (Gonzalez-Ferreras et al., 2012):

- Tilt has its origin in the need to represent the relevant movements of the pitch contours in text-to-speech applications (Taylor, 2009). The Tilt parameters are obtained by a combination of RFC parameters, with a numerical approximation of the prosodic events. (Ananthakrishnan and Narayanan, 2008a; Rosenberg, 2009).
- Bézier parameters are obtained as an approximation of the pitch contours with Bézier functions (Escudero and Cardenoso, 2007). The minimum square fitting approximation technique is used to represent the shape of the F0 contour along a given reference unit. The control points of the spline are the parameters that project the temporal evolution of the pitch contour. In this work, we use 4 points as Bézier parameters. The Bézier approximation has similarities with other proposals such as quantified contour modeling (Rosenberg, 2010). Both proposals have the advantage of allowing an increase in the number of parameters, in terms of the required accuracy.

The use of context features can significantly improve the classification results, as has been previously reported (Gonzalez-Ferreras et al., 2012; Levow, 2005; Rosenberg and Hirschberg, 2009). The inclusion of all of the features from the previous and following words will result in too many features. Thus, we decided to select the features to model the context using the Correlation-based Feature Selection (CFS) algorithm (Hall, 1998). We used 18 features for each word, without the use of context. The CFS algorithm selected 8 features to model the context, as described in (Gonzalez-Ferreras et al., 2012). We tried different configurations, and the best configuration was using 2 previous words and 2 following words as context. The experiments reported in this paper used this configuration.

3. The automatic fuzzy labeling tool operation

As shown in (Gonzalez-Ferreras et al., 2012), binary classifiers provide high accuracy labeling results. Thus, we propose to combine binary classifiers using *pairwise coupling* in order to achieve better accuracy results in the complex multi-class automatic labelling problem. Three different classifiers (neural networks, decission trees and support vector machines) are used to provide scores to be combined in the final decission stage, where the outputs of these three classifier modules are combined using the comprehensive fuzzy technique proposed in (Escudero-Mancebo et al., 2014). As a result, the fuzzy classifier provides a list of candidate labels for each word, and each label has a numerical certainty level attached. Precise details about specific aspects of the fuzzy labeling technique are included in (Escudero-Mancebo et al., 2014). Here, we summarize the essential parts of that technique to help understanding the approach followed in this paper.

Given an input vector $\mathbf{x}$ and a set of labels $\mathcal{L} = \{l_1 \dots l_C\}$, classic pattern recognition assigns a unique label l^* to $\mathbf{x}$. The classification rule selects the label l^* which maximizes the posterior probability:

$$l^* = \arg\max_{l} P(l|\mathbf{x}) \tag{1}$$

The objective of fuzzy classification is to obtain membership values μ_i as an estimation of $P(l_i|\mathbf{x})$. This vector of membership values $\boldsymbol{\mu} = \mu_1 \dots \mu_i \dots \mu_C$ is used in the decision making process.

The pairwise coupled approach basically divides a given multiclass classification problem into a number of binary classification sub-problems, from which the results must be combined to obtain the final classification result (Hastie and Tibshirani, 1998; Wu et al., 2004). According to this approach, let us refer by $\hat{P}(l_i|\mathbf{x}, \lambda^k_{l_i,l_j})$ to an estimation of the probability $P(y = l_i|\mathbf{x}, y = l_i \vee l_j)$, where l_i and l_j are two different prosodic labels; $\mathbf{x}$ is the input of the classifier (the prosodic features); y is the class label; and $\lambda^k_{l_i,l_j}$ is a pairwise classifier of type k (in our case, $k = 1$ for a neural network, $k = 2$ for a decision tree, or $k = 3$ for a support vector machine) that is trained to separate classes l_i and l_j. There are as many classifiers as there are combinations of pairs of C classes: $\frac{C\cdot(C-1)}{2}$. Each classifier, $\lambda^k_{l_i,l_j}$, provides the posterior probability estimations $\hat{P}(l_i|\mathbf{x}, \lambda^k_{l_i,l_j})$ and $\hat{P}(l_j|\mathbf{x}, \lambda^k_{l_i,l_j})$.

In order to combine the results of the pairwise classifiers of type k, we followed the approach described in (Hastie and Tibshirani, 1998), which generates multiclass probability estimates from a combination of outputs of the pairwise binary classifiers. The iterative algorithm tries to minimize the Kullback-Leibler distance between the estimation of the probablity $P(l_i|\mathbf{x}, \lambda^k_{l_i,l_j})$ and $\hat{P}(l_j|\mathbf{x}, \lambda^k_{l_i,l_j})$ and the values jointly predicted for it by the different classifiers. The algorithm provides an estimation of the probabilities $P(l_i|\mathbf{x}, \lambda^k)$ for each classifier $\lambda^1 \dots \lambda^k \dots \lambda^M$, and for each class $l_1 \dots l_i \dots l_C$.

Thus, each classifier λ^k takes the vector $\mathbf{x} \in \mathcal{R}$ as the input and generates a C-dimensional vector at the output:

$$\boldsymbol{\mu}^{\mathbf{k}}(\mathbf{x}) = [\mu^k_1(\mathbf{x}), ..., \mu^k_C(\mathbf{x})] \quad (2)$$

where $\mu^k_i(\mathbf{x})$ is the degree of support from the classifier k to the hypothesis that the instance $\mathbf{x}$ comes from the class l_i. The outputs of the M classifiers can be expressed in a *decision profile* (Kuncheva et al., 2001):

$$DP(\mathbf{x}) = \begin{bmatrix} \mu^1_1(\mathbf{x}) & .... & \mu^1_i(\mathbf{x}) & ... & \mu^1_C(\mathbf{x}) \\ ... & ... & ... & ... & ... \\ \mu^k_1(\mathbf{x}) & .... & \mu^k_i(\mathbf{x}) & ... & \mu^k_C(\mathbf{x}) \\ ... & ... & ... & ... & ... \\ \mu^M_1(\mathbf{x}) & .... & \mu^M_i(\mathbf{x}) & ... & \mu^M_C(\mathbf{x}) \end{bmatrix} \quad (3)$$

The next step is to combine the results of the M independent classifiers. In this work we use the *product combination* and the *fuzzy integral* techniques.

The product combination (Kittler et al., 1998) calculates the support for class l_i using the i^{th} column of $DP(\mathbf{x})$:

$$\mu_i(\mathbf{x}) = \prod_{k=1}^{M} \mu^k_i(\mathbf{x}) \quad (4)$$

The *fuzzy integral* technique is described in (Grabisch, 1995; Grabisch and Sugeno, 1992) and has been used to combine classifiers in several contexts (Benediktsson et al., 1997; Cho and Kim, 1995a; Cho and Kim, 1995b; Gader et al., 1996; Verikas et al., 1999; Wang et al., 1998). The support for the class l_i, $\mu_i(\mathbf{x})$ is the compromise between the *competence* of the classifier represented by the measure g and the *evidence* represented by the i^{th} column of the decision profile $DP(\mathbf{x})$. When the fuzzy integral algorithm is applied, a fuzzy set is obtained $\boldsymbol{\mu}(\mathbf{x}) = \mu_1(\mathbf{x}), ..., \mu_C(\mathbf{x})$, μ_i being an estimation of $P(l_i|\mathbf{x})$.

Finally, after applying the α-cuts algorithm (see (Escudero-Mancebo et al., 2014) for details), we have $\boldsymbol{\mu_\alpha}(\mathbf{x}) = \mu_{1_\alpha}(\mathbf{x}), ..., \mu_{C_\alpha}(\mathbf{x})$, so that $\mu_{i_\alpha} \in \{0, 1\}$, and three situations can occur:

1. $\mu_{i_\alpha}(\mathbf{x}) = 0 \; \forall i \in 1..C$. The classifier is not able to assign any label to $\mathbf{x}$ because the evidence is not high enough.
2. $\exists i$ so that $\mu_{i_\alpha}(\mathbf{x}) = 1$ and $\mu_{j_\alpha}(\mathbf{x}) = 0 \; \forall j \in 1..C, j \neq i$. The classifier assigns only one label to $\mathbf{x}$.
3. The rest of the cases in which the fuzzy classifier interprets that more than one label could be assigned to the input prosodic unit $\mathbf{x}$.

The classifier fails when the unit $\mathbf{x}$ has been labeled as l in the testing corpus, but $\mu_{l_\alpha}(\mathbf{x}) = 0$. The results presented in the following sections are interpreted taking into account the three possible situations.

4. Results

4.1. Classification recall

Table 3 (upper sub-table) shows the different labelers' automatic classification recall to identify the categories of boundary tones. The description of the pitch accents and breaks labeling performance is omitted in this paper due to the lack of space. Some of the classifiers seem to be specialized in the identification of certain labels (e. g. the SVM classifier outperforms the rest of classifiers in the identification of the L% boundaries: 88.8% vs. 81.9% and 84.3%). The product fusion strategy (Gonzalez-Ferreras et al., 2012) benefits from this fact and results in a better overall performance: 81.2% vs 79.9%, 76.4% and 78.3%

Table 3 (lower sub-table) compares the results obtained using the product fusion strategy with ones obtained with the fuzzy integer strategy. The recall metric used in the fuzzy classifiers differs from the one used in the non-fuzzy classifiers. Fuzzy classifiers can assign more than one label to each word so that in positive cases the correct label belongs to the predicted set of labels. On the other hand, a negative case implies that the correct label is not in the set of predicted labels. As can be seen in the table, the proper selection of the α-value permits to obtain a performance clearly higher than the one obtained in the non-fuzzy versions. The tuning of the α-value is a crucial task as it has important implications both in the classification recall and in the number of predicted labels per word.

In the right side of the Figure 1, $\alpha = 0$ implies *Total Recall = 100%*; in the left side of the Figure, $\alpha = 1$ implies *Total Recall = 0%*. The reason is that $\alpha = 0$ (null certainty) implies that all the labels are predicted. In between these extreme points, a compromise can be found between the recall and the mean number of predicted labels. The red line of the plot (*% Fails*), has a local maximum at α_{max} that has to be avoided. We select in this case $\alpha = 0.38 < \alpha_{max}$ because the recall increases at the time that the *% Empty output* is zero. Decreasing the number of empty predictions is important in our scenario because we implement a system to support manual transcribers who check the automatic predictions and we estimate that the revision of the predicted labels is less expensive than coding.

Table 4 analyzes the behavior of the automatic labeler in the neighborhood of the local maximum of the curve *% of Fails* in Figure 1. The classifier predicts only one label per word most of the time (75%). The amount of words without

Classifer	L%	H%	=%	!H%	LH%	none	Total
Decision Tree (DT)	84.3%	54.3%	12.5%	25.5%	34.7%	92.2%	78.3%
Support Vector Machine (SVM)	81.9%	55.2%	29.5%	41.4%	40.3%	86.7%	76.4%
Neural Network (NN)	88.8%	56.6%	17.9%	28.5%	40.3%	92.9%	79.9%
Product Fusion	88.4%	61.3%	14.3%	38.5%	41.7%	93.1%	81.2%

Classifer	L%	H%	=%	!H%	LH%	none	Total
Product Fusion	88.4%	61.3%	14.3%	38.5%	41.7%	93.1%	81.2%
Fuzzy Integer Fusion α=0.5	83.5%	44.8%	3.6%	17.2%	27.8%	90.3%	74.8%
Fuzzy Integer Fusion α=0.35	95.2%	77.0%	39.3%	57.7%	55.6%	97.1%	88.8%
Fuzzy Integer Fusion α=0.32	96.0%	80.4%	44.6%	61.5%	63.9%	97.8%	90.4%
Fuzzy Integer Fusion α=0.2	96.8%	88.5%	58.9%	78.7%	76.4%	99.2%	94.4%

Table 3: Recall of the boundary tone labeling task using different classifiers and different combination techniques

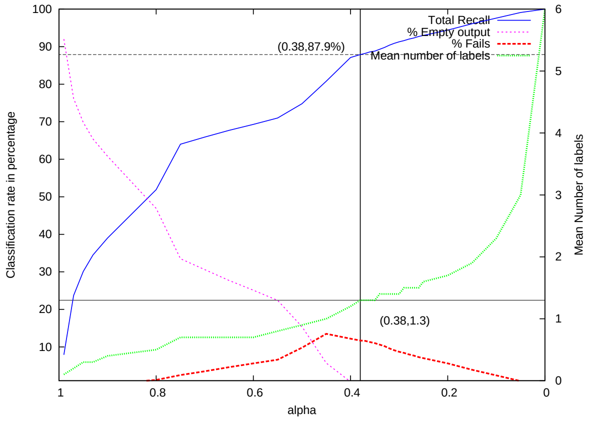

Figure 1: Evolution of total classification rate and mean number of labels in terms of α in the boundary tone labeling task.

predictions is close to zero. Only 2% of the predictions has more than two labels per word. When the classifier predicts two or more labels per word, the recall is very high (only 4% of incorrect predictions). This rates gives the manual transcriber a high degree of confidence in the presence of the real label inside the predicted set of labels, speeding up the revision process.

4.2. Decision template

The predictions of the fuzzy classifier are visualized in a multi-tier window, aligned with the corresponding prosodic event (pitch accents aligned with the stressed syllables, tone boundaries aligned with the end of the word) (see figure 2). The advantage of the fuzzy classifier when compared with conventional classifiers is that it can provide more than one label per prosodic unit as a function of the uncertainty of the predictions. Each tag is accompanied by a numeric value in the [0,1] interval: a higher value represents bigger certainty. At this point, it should be noted that the procedure, according to the fuzzy sets theory, is not based on probabilities since the degrees of certainty can sum up to more than 1: the fact that three labels are proposed is not equivalent to say that they represent three alternatives summing for a probability 1, since the degree of certainty is independently assigned for each category. This explains the fact that even when only one tag is predicted, it is not necessarily accompanied by complete confidence (marked with 1). On the other hand, having more than one tag in the output represents a difficult situation in which more than one label evidences a degree of certainty over the threshold set by the α-cut.

It can be seen in figure 2 that each word is labeled with a different number of tags. Each of the tags is associated with a degree of certainty in the [0,1] interval and tags are ordered from the highest to the lowest degree. In the example, the most certain decision is the label “none” (absence of pitch accent) assigned to the word “las” (the): only one label is assigned and the degree of certainty is 0.8. On the other hand, the most uncertain decision corresponds to the pitch accent of the word “veladas” ($evenings$), where three different labels are proposed as candidates. Moreover, the figures assigned to this word go from 0.5 to 0.34, which are significantly lower than the 0.8 assigned to the word “las” (the).

The fuzzy classifier results were checked by a human transcriber. She reviews the template generated by the system and selects the right choice on this template. If none of the candidate labels is the right one, the transcriber has to enter it. The example in figure 2 illustrates three different type of situations that the transcriber has to deal with. Firstly, in the case of the pitch accent associated to the stressed syllable of the word “volumen” ($edition$), the fuzzy classifier offers a unique solution L+H*, but with a low degree of certainty, 0.51. In this case, the human labeler checks if the solution is the right one according to his/her perception and marks it with (+). If that is not the case, the transcriber writes a new option. Secondly, concerning the boundary tone associated to the final syllable of the word “volumen” (edition), the fuzzy classifier proposes three candidates: the absence of break (none (0.39)) and the choice between two boundary tones that are conceptually similar, H% (0,43) and !H% (0.43). This example shows the interaction between the break level and the tone level in the ToBI system. Once the labeler decides that there is a prosodic break, he/she has to decide if the boundary tone is high (H%) or mid (!H%).

#labels	$\alpha = 0.65$		$\alpha = 0.50$		$\alpha = 0.35$		$\alpha = 0.38$		$\alpha = 0.20$	
	incorr.	correct	incorr.	correct	incorr.	correct	incorr.	correct	incorr.	correct
0	28%	0%	15%	0%	0%	0%	0%	0%	0%	0%
1	5%	68%	10%	74%	6%	66%	7%	68%	2%	55%
2	0%	0%	2%	1%	5%	19%	4%	18%	2%	20%
3	0%	0%	0%	0%	0%	4%	0%	2%	2%	15%
4	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%
Total	33%	68%	25%	75%	11%	89%	11%	88%	6%	94%

Table 4: Rates of correctly classified samples (boundary tone labeling task) and incorrectly classified samples in terms of the number of labels assigned using the fuzzy classifier with different values of α.

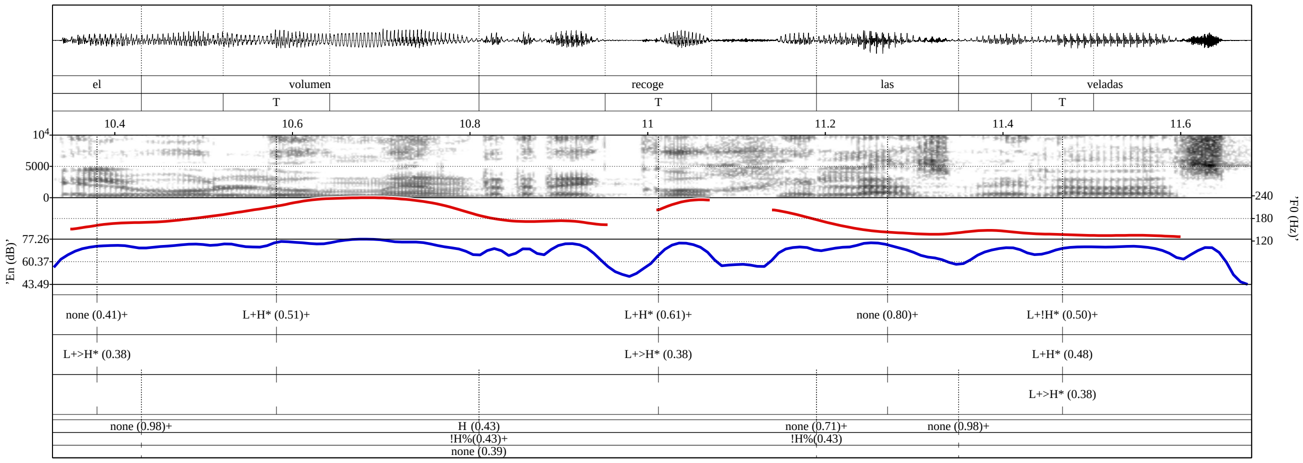

Figure 2: Multi-tier prediction interface that allows a human labeler to check the output of the automatic fuzzy classifier. The reviewer selects the label that she considers the correct one from a set of limited alternatives. In this sample utterance, the predicted labels have been generated by the fuzzy classifier. Word and syllable segmentation appear in the upper tiers. (T means tonic syllable). The six bottom tiers correspond to the predicted pitch accent and boundary tones labels. In parenthesis is the degree of support of the decision resulting from the fuzzy classifier. The symbol "+" after the parenthesis is inserted by the reviewer for marking the correct label.

Acoustically, there are no clear differences between the nuclear accent associated to the word "volumen" and its final accent, but since a decrease of the tonal range is perceived at the end of the word, the transcriber selects the second option, that is, !H% (0.43). Thirdly, the fuzzy classifier proposes a tone with differences in alignment (which has been demonstrated as phonologically relevant in Spanish, see (Face and Prieto, 2007)) for the pitch accent associated to the stressed syllable of the word "recoge" ($collects$): L+H* (0.61) and L+>H* (0.38). The transcriber rejects the displaced tone since the peak is within the stressed syllable. Consequently, the first option, with a higher degree of certainty L+H* (0.61) is marked with a + sign.

4.3. Performance of the template revision

Tuning of the α-cut value is crucial for a correct system operation. The lower the value of the α-cut, the higher the number of positive cases (a positive case occurs when the right label is in the set of predicted labels, computed as the *soft-classification rate* in (Escudero-Mancebo et al., 2014)). For the reviewer, it is important to know that the probability of having the real label in the set of candidates is really high because the number of corrections will be potentially low. But, on the other hand, increasing the number of labels will make the selection of the correct one harder. In our case, in the training stage we obtained soft classification rates of 82% for pitch accents and 88% for boundary tones. These rates are clearly higher than the accuracy rates that we obtain in classic non-fuzzy classification (69.2% and 81.2% respectively). This increase in classification rates is expected to improve the performance of the reviewing process.

First results report that in most cases (81.8% for boundary tones and 72.6% for pitch accents) the labeler's option is the first candidate. Only 9.2% of the boundary tones needed and 13.5% of the pitch accents labels needed to be edited. The consequence is that the labelling speed is increased. In a preliminary test, the real time ratio (Syrdal et al., 2001) of the labeling process was 1:66 when the template of predictions is used. This ratio contrasts with the one obtained without any supporting template which was 1:80.

The multi-label approach is a useful method to represent the uncertainty that characterizes the prosodic labeling. It is again demonstrated (as in (Escudero-Mancebo et al., 2014) when the methodology has been applied to the Boston Radio News Corpus) that the label candidates proposed by the fuzzy classifier are the ones on which human labelers do not usually agree. In (Escudero-Mancebo et al., 2014) can be found a detailed analysis of these cases.

5. Conclusions and future work

In this paper, an automatic prosodic labeling system is described. The system is based on the use of pairwise classification and fuzzy integral expert fusion aggregation techniques. The system has been trained and tested for predicting SP_ToBI labels in the Glissando corpus. The novelty with respect to other approaches is that the system predicts more than one label per word. A human transcriber is responsible for the selection of the correct label.

We have evaluated the performance of the automatic labeler by comparing the labeling speed of a human transcriber when she uses the predictions of the system or not, obtaining encouraging results.

The predicted labels are annotated with a degree of certainty that represents the confidence of the system on assigning the label. This certainty degree mimics in some way the doubts that human transcribers also manifest in the annotation processes. Indeed, the most confused pair of labels that appear in the inter-transcriber tests are also the most frequent pair of labels in the revision template.

The revisions of the automatic labels are used to retrain the system in order to obtain more accurate predictions. Our present concern is the definition of appropriate metrics that permit to contrast the goodness of the reviewing-retraining iterative process.

A friendly interface is crucial on the improvement of the revision performance. We profit the revision process to obtain feedback from the transcriber that permits to improve the revision template interface.

The Sp_ToBI labels (both the manual and the automatic ones) are freely available in the web page of the Glissando project[1]. And we expect to be distributed in the framework of the Reciprosody project[2].

6. References

S. Ananthakrishnan and S. Narayanan. 2008a. Fine-grained pitch accent and boundary tone labeling with parametric F0 features. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pages 4545–4549.

S. Ananthakrishnan and S.S. Narayanan. 2008b. Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):216–228, January.

M.E. Beckman. 2002. Intonation across Spanish in the Tones and Break Indices framework. *Probus*, 14:9–36, January.

J. A. Benediktsson, J. R. Sveinsson, J. I. Ingimundarson, H. Sigurdsson, and O. K. Ersoy. 1997. Multistage classifiers optimized by neural networks and genetic algorithms. *Nonlinear Anal., Theory, Meth., Applicat.*, 30(3):1323–1334.

S.-B. Cho and J. H. Kim. 1995a. Combining multiple neural networks by fuzzy integral and robust classification. *IEEE Trans. Syst., Man, Cybern.*, 25:380–384.

S. B. Cho and J. H. Kim. 1995b. Multiple network fusion using fuzzy logic. *IEEE Trans. Neural Networks*, 6:497–501.

G. Elordieta. 2011. Transcription of intonation of the Spanish language. In *Estudios de Fonética Experimental*, volume XX, pages 273–293.

David Escudero and Valentin Cardenoso. 2007. Applying data mining techniques to corpus based prosodic modeling. *Speech Communication*, 49(3):213–229.

D. Escudero, V. Cardeñoso, and A. Bonafonte. 2002. Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish. In *Proceedings of ICASSP 2002*, volume 1, pages 481–484, May.

David Escudero, Lourdes Aguilar, Maria del Mar Vanrell, and Pilar Prieto. 2012. Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system. *Speech Commun.*, 54(4):566–582, May.

David Escudero, Lourdes Aguilar, César González, Valentín Cardeñoso, and Yurena Gutiérrez. 2014. Preliminary results on Sp_ToBI prosodic labeling assisted by an automatic fuzzy classifier. In *Proceedings of International Conference on Speech Prosody*, page in press, Dublin, Ireland, May.

David Escudero-Mancebo and Eva Estebas-Vilaplana. 2012. Visualizing tool for evaluating inter-label similarity in prosodic labeling experiments. In *Proceedings Interspeech 2012*.

David Escudero-Mancebo, Cesar González-Ferreras, Carlos Vivaracho-Pascual, and Valentin Cardeñoso Payo. 2014. A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling. *Computer Speech and Language*, 28(1):326 – 341.

E. Estebas Vilaplana and P. Prieto. 2009. La notación prosódica en español. una revisión del sp_tobi. *Estudios de Fonética Experimental*, XVIII:263–283.

E. Estebas Vilaplana and P. Prieto. 2010. Castilian Spanish Intonation. In P. Prieto and P. Roseano, editors, *Transcription of Intonation of the Spanish Language*, pages 17–48. Lincom Europa, München.

T. Face and P. Prieto. 2007. Rising accents in Castilian Spanish: a revision of Sp-ToBI. *Journal of Portuguese Linguistics*, 6.1:117–146.

P. D. Gader, M. A. Mohamed, and J. M. Keller. 1996. Fusion of handwritten word classifiers. *Pattern Recogn. Lett.*, 17:577–584.

Juan-María Garrido, David Escudero, Lourdes Aguilar, Valentín Cardeñoso, Emma Rodero, Carme de-la Mota, César González, Carlos Vivaracho, Sílvia Rustullet, Olatz Larrea, Yesika Laplaza, Francisco Vizcaíno, Eva Estebas, Mercedes Cabrera, and Antonio Bonafonte. 2013. Glissando: a corpus for multidisciplinary prosodic

[1] http://veus.glicom.upf.edu/

[2] http://jaguar.cs.qc.cuny.edu/

studies in Spanish and Catalan. *Language Resources and Evaluation*, 47(4):945–971.

C. Gonzalez-Ferreras, D. Escudero-Mancebo, C. Vivaracho-Pascual, and V. Cardeñoso Payo. 2012. Improving automatic classification of prosodic events by pairwise coupling. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(7):2045 –2058, sept.

M. Grabisch and M. Sugeno. 1992. Multi-attribute classification using fuzzy integral. In *IEEE Int. Conf. Fuzzy Systems*, pages 47–54.

M. Grabisch. 1995. On equivalence classes of fuzzy connectives – the case of fuzzy integrals. *IEEE Trans. Fuzzy Syst.*, 3:96–109.

M. A. Hall. 1998. *Correlation-based Feature Subset Selection for Machine Learning*. Ph.D. thesis, University of Waikato, Hamilton, New Zealand.

Trevor Hastie and Robert Tibshirani. 1998. Classification by pairwise coupling. *The annals of Statistics*, 26(2):451–471, April.

J. I. Hualde. 2003. El modelo métrico autosegmental. In P. Prieto, editor, *Teorías de la entonación*, pages 155–184. Ariel.

J. Kittler, M. Hatef, R. P W Duin, and J. Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March.

L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin. 2001. Decision templates for multiple classifier fusion: An experimental comparison. *Pattern Recogn.*, 34(2):299–314.

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

Kikuo Maekawa. 2004. Design, compilation, and some preliminary analyses of the corpus of spontaneous japanese. *Spontaneous speech: Data and analysis*, 3:87–108.

T. Navarro-Tomás. 1944. *Manual de Entonación Española*. Madrid, Guadarrama.

M. Ostendorf, P.J. Price, and S. Shattuck. 1995. The Boston University Radio News Corpus. Technical report, Boston University.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

J. F. Pitrelli, M. E. Beckman, and J. Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of ICSLP*, pages 123–126.

M.A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymon. 2005. 45:89–95.

P. Prieto and P. Roseano, editors. 2010. *Transcription of Intonation of the Spanish Language*. Lincom Europa, München.

V.K. Rangarajan Sridhar, S. Bangalore, and S.S. Narayanan. 2008. Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):797–811, May.

Andrew Rosenberg and Julia Hirschberg. 2009. Detecting Pitch Accent at the Word, Syllable and Vowel Level. In *HLT/NAACL*, pages 81–84.

A. Rosenberg. 2009. *Automatic Detection and Classification of Prosodic Events*. Ph.D. thesis, University of Columbia, USA.

A. Rosenberg. 2010. Classification of Prosodic Events using Quantized Contour Modeling. In *HLT/NAACL*, pages 721–724.

J. M. Sosa. 2003. La notación tonal del español en el modelo Sp-ToBI. In P. Prieto, editor, *Teorías de la entonación*, pages 155–184. Ariel.

A. Syrdal and J. McGory. 2000. Inter-transcriber reliability of ToBI prosodic labeling. In *Proceedings of ICSLP*, volume 3, pages 235–238.

A. K. Syrdal, J. Hirshberg, J. McGory, and M. Beckman. 2001. Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, (33):135–151.

P. Taylor. 2000. Analysis and synthesis of intonation using the Tilt model. *Journal of Acoustical Society of America*, 107(3):1697–1714.

Paul Taylor. 2009. *Text-to-Speech Synthesis*. Cambridge University Press.

A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, and A. Gelzinis. 1999. Soft combination of neural classifiers: A comparative study. *Pattern Recogn. Lett.*, 20:429–444.

D. Wang, J. M. Keller, C. A. Carson, K. K. McAdoo-Edwards, and C. W. Bailey. 1998. Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion. *IEEE Trans. Syst., Man, Cybern.*, 28B:583–591.

Colin W Wightman. 2002. Tobi or not tobi? In *Speech Prosody 2002, International Conference*.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, December.

T.J. Yoon, S. Chavarría, J Cole, and M. Hasegawa-Johnson. 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of Interspeech, Jeju*, pages 2729–2732.