Improving automatic classification of prosodic events by pairwise coupling

César González-Ferreras*, David Escudero-Mancebo, Carlos Vivaracho-Pascual and Valentín Cardeñoso-Payo, *Member, IEEE*

***Abstract*—This paper presents a system that automatically labels ToBI events. The detection (binary classification) of prosodic events has received significantly more attention from researchers than its classification because of the intrinsic difficulty of classification. We focus on the classification problem, identifying 8 types of pitch accent tones, 9 types of boundary tones and 5 types of break indices. The complex multi-class classification problem is divided into several simpler problems, by means of pairwise coupling. We propose to combine two-class classifiers to achieve the multi-class classification because two-class problems provide high accuracy results. Furthermore, complementarity between Artificial Neural Networks and Decision Trees classifiers has been exploited to improve the final system, combining their outputs using a fusion method. This proposal, together with the adequate feature extraction that includes the use of features such as the Tilt and Bézier parameters, allows us to achieve a total classification accuracy of 70.8% for pitch accents, 84.2% for boundary tones and 74.6% for break indices, on the Boston University Radio News Corpus. The analysis of the misclassified samples shows that the types of mistakes that the system makes do not differ significantly from the common confusions that are observed in manual ToBI inter-transcriber tests.**

***Index Terms*—ToBI labeling, prosodic event classification, pairwise classifiers, classifiers fusion, spoken language processing**

I. Introduction

The tones and break indices (ToBI) is a system for labeling prosodic events that are perceived in spoken utterances [1], [2]. This standard distinguishes three types of prosodic events: Pitch Accents, Boundary Tones and Break Indices. Pitch Accents refer to the prosodic function of prominence, and they are set with a combination of two basic tones: an *H* (high tone) and an *L* (low tone). Boundary Tones and Break Indices refer to the prosodic function of phrasing. Boundary Tones use the same two basic symbols, *H* and *L*, to label the intonational boundaries. Break Indices mark the degree of separation of the words. The *detection* of these prosodic events has received significantly more research attention than its *classification* [3] because of the greater difficulty of the latter. This paper focuses on the *labeling* of ToBI events, for the problem of detecting and classifying them, which means distinguishing 8 types of Pitch Accents, 9 types of Boundary Tones and 5 types of Break Indices, which are defined in the standard.

The labeling of prosodic events can be very useful for practical applications, for example, the detection of speech acts, the disambiguation of words and the improvement of the performance of speech recognition systems and text-to-speech systems [4]. Increasing the granularity of the presence or absence of ToBI event decisions can be important because the ToBI symbol sequences can be associated with a meaning or with a prosodic function that is projected in the acoustic characteristics of the utterance. Thus, for example, the Pitch Accent *L+H** is associated with a *contrastive focus* at the time that *L** refers to *yes-no questions* [5]. The extraction of this information can be useful for the understanding of the message. Furthermore, the modeling of the relationship between such sequences of tones and the corresponding prosodic shape would be very useful for increasing the naturalness of text-to-speech applications [6]. Another more straightforward application is to increase the speed of performing manual ToBI tagging of corpora, which offers an automatic label proposal for a manual transcriber to revise. The manual prosodic labeling is estimated to take 100-200 times real time [7].

Until the time that the detection of prosodic events is addressed with efficiency and becomes part of the state-of-the-art (Section I-A reviews the state of the art, with accuracy rates close to 90%), the classification problem will be considered to be a difficult task that is simplified by grouping symbols [3] or reducing the number of speakers [6], [8]. In [9], we discuss three of the main factors that make this problem difficult: the lack of a definitive parameterization technique of the intonation contours, the high inter ToBI symbol similarity and the imbalanced nature of the training labeling corpora. In this paper, we propose a methodology for the classification of prosodic events that is independent of the input features, taking into account the imbalanced nature of the prosodic corpora and facing inter-class similarity with specialized classifiers.

The use of a set of appropriate input features is the key to obtaining good classification results. In prosodic labeling, the F0, energy and duration play an important role for marking both the emphatic and phrasing prosodic functions [10]. In [4], a set of statistical variables that measure the variation of F0, energy and duration in the syllables are computed for the detection of the ToBI events (see section II for a review of such input features). For classifying events, it is important to measure the temporal evolution of the intonational contours

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) and by Consejería de Educación de la Junta de Castilla y León (project VA322A11-2).

The authors are with the Department of Computer Science, Universidad de Valladolid, Campus Miguel Delibes, 47011 Valladolid, Spain (e-mail: cesargf@infor.uva.es; descuder@infor.uva.es; cevp@infor.uva.es; valen@infor.uva.es).

because there is a correspondence between the ToBI symbols and the shape of the contour (e.g., the *H%* pattern is an ascending contour and the *L%* pattern is descending). The shape of the F0 contour has been used in other state-of-the-art approaches: Tilt parameters [11] are used in [8], quantized contours are used in [3] and stylized contours are used in [9]. This paper shows that the inclusion of additional specific input features that model the temporal evolution of the F0 contour (Tilt and Bézier parameters) helps to better distinguish between different types of pitch accents and boundary tones.

The state-of-the-art offers very few ToBI tagged corpora. The Boston University Radio News Corpus (BURNC) [12] is the main reference. This corpus is clearly imbalanced with respect to what concerns the presence of different ToBI symbols (for example, the *H** class has more than ten times the number of samples than the *L** class). Furthermore, this imbalance does not only depend on the corpus, since the differences in relative frequency of labels depend on the language, as it is well known. Although some data sampling techniques were systematically analyzed in [13] to try to reduce this impact in pitch accent type classification tasks, it is a fact that the classification performance is negatively affected. On the other hand, we showed in previous work that, while a decision tree became specialized in the most populated classes, a multilayer perceptron balanced the results of the classification of the different classes better, independently of the size of the corpora [9]. In this work, we configure a specific architecture that fuses the results of different types of classifiers, improving the performance on the overall classification task.

The manual ToBI labeling of corpora is commonly performed by teams of evaluators who are trained to follow a common guideline [5]. In spite of the rigorous quality controls that are applied to this process, inconsistencies appear. Even the Boston University Radio News Corpus, for which inter-transcriber agreement rates are higher than 90%, also reflects this fact, even when including comments of the labelers concerning their doubts about the judgments [12]. The reason for this outcome is the high interclass similarity of some of the symbols. Reference [14] reports on the similarity of the ToBI symbols when taken pairwise. Their conclusions are supported by empirical inter-transcriber judgments and the opinions of the labelers about the conceptual similarity of every pair of symbols. The inter-pair similarity of the symbols was the reason for the decision to merge symbols in some of the studies on the classification of ToBI events, such as [6], [3] or [8]. Our alternative hypothesis is that the use of pairwise coupled classifiers can be efficient when tackling interclass similarity, which we have empirically observed [9]. The distinction of classes in pairs is a relative easy task, but the multi-class classification problem reduces the performance dramatically. Because two-class problems have better performance, we propose to combine two-class classifiers for a multi-class classification. As a result, the error rate decreases in comparison with other state-of-the-art works, and the most often confused pairs of classes categorized by the classifier are the same as the pairs for which manual transcribers previously disagreed.

A. State of the art

Pioneering studies in this field of research date back to the 1990s, with the contributions of Prof. Ostendorf et al. in [18] and [6]. In [18], they present a strategy that combines Decision Trees and Markov Models for detecting accents, boundaries and breaks in the BURNC corpus. Decision trees provide the probability of each class and Markov sequence models provide the probability of the sequence of classes, which is implemented with a Viterbi algorithm. In [6], they predict ToBI labels from text with decision trees and Markov sequence models that have been previously trained with the samples of one of the speakers of the BURNC corpus. The duration, energy and F0 features are used to train the automatic prediction models in both cases. Some similar accent types are grouped in the same class. The reduced set of symbols is used to predict the F0 contours in [19] in the context of speech synthesis. With the same reduced set, [17] slightly improves the results by including bagging and boosting techniques in the decision tree learning strategy. Tables I, II and III report on the classification rates that are obtained and on the ToBI symbols that are used.

More recently, the works of Prof. Narayanan et al. focused on the issue again. First, the maximum entropy model was used with acoustic (F0, energy and duration), lexical and syntactic features. The proposed framework labels the pitch accent, the boundary tone and the prosodic break index at the word level [15]. Second, a system that labels the pitch accent and the boundary tone at the syllable level is reported [4]. A different model is used for the acoustic features, the lexical features and the syntactic features: the acoustic model combines a classifier (Linear Discriminant, Gaussian Mixture Models or Artificial Neural Networks) with n-grams; the lexical and syntactic features are modeled using n-grams. Then, the three models are combined. Both of the studies focused on the binary detection problem. More experiments were performed, but they were performed on fine-grained pitch accents and boundary tone labeling, using the Tilt parameters with n-grams [8] and Hidden Markov models [20]. They also used the BURNC corpus and reduced the number of ToBI labels by grouping them. Tables I, II and III report on the results of these studies.

The studies that focus exclusively on the detection of the ToBI tones and breaks problem are not taken into account in this review. An excellent revision about this concern can be found in [4]. In 2009, Rosenberg defended a doctoral thesis on automatic detection and classification of prosodic events in the laboratory of Prof. Hirschberg [13], whose results evolved into the AuToBI freeware tool for the ToBI labeling of spoken corpora [21]. They quantify the F0 contours and model the context where the ToBI symbol is supposed to be produced [3]. A simplified version of ToBI is also used, and the results that are reported in [3] are summarized in Tables I and II. In [22], the presence of a pitch accent is predicted from the text, and in [7], an automatic proposal of prosodic events is generated, as a method for speeding up the manual labeling of a given corpus.

The interdisciplinary group on prosodic studies of the

TABLE I
ACCURACY OF THE PITCH ACCENT TONE CLASSIFICATION FOR DIFFERENT MAPPINGS OF THE TOBI LABELS: A COMPARISON BETWEEN THE STATE OF THE ART AND OUR APPROACH. BOSTON UNIVERSITY RADIO NEWS CORPUS HAS BEEN USED IN ALL CASES.

		Detection		Classification				
Mapping	H*	Accent	Accent	H*	H*	High	High	High
	L+H*	Accent	Accent	L+H*	L+H*	High	High	High
	!H*	Accent	Accent	H*	!H*	Downstepped	Downstepped	Downstepped
	H+!H*	Accent	Accent	H+!H*	ignored	High	High	High
	L+!H*	Accent	Accent	L+H*	ignored	Downstepped	Downstepped	Downstepped
	L*	Accent	Accent	L*	L*	Low	Low	Low
	L*+H	Accent	Accent	L*+H	ignored	Low	Low	Low
	no label	none	none	ignored	ignored	Unaccented	Unaccented	Unaccented
	#Classes	2	2	5	4	4	4	4
State of the Art	Reference	[15]	[4]	[3]	[8]	[6]	[16]	[17]
	Level	word	syllable	word	word	syllable	syllable	syllable
	#Words/Syllables	24,955	44,390	29,578	28,300	14,599	14,599	14,377
	#Speakers	4	6	6	6	1	1	1
	Accuracy	**86.0%**	**86.75%**	**63.99%**	**56.4%**	**80.17%**	**81.3%**	**87.17%**
This Work	Level	word		word	word	word		
	#Words	27,767		27,767	27,767	27,767		
	#Speakers	6		6	6	6		
	Accuracy	**86.7%**		**69.1%**	**63.9%**	**80.0%**		

TABLE II
ACCURACY OF THE BOUNDARY TONE CLASSIFICATION FOR DIFFERENT MAPPINGS OF THE TOBI LABELS: A COMPARISON BETWEEN THE STATE OF THE ART AND OUR APPROACH. BOSTON UNIVERSITY RADIO NEWS CORPUS HAS BEEN USED IN ALL CASES.

		Detection		Classification		
Mapping	L-L%	btone	btone	L-L%	L-L%	L-L%
	!H-L%	btone	btone	!H-L%	ignored	ignored
	H-L%	btone	btone	H-L%	ignored	H-L%
	L-H%	btone	btone	L-H%	L-H%	L-H%
	H-H%	btone	btone	H-H%	ignored	ignored
	L-	btone	btone	ignored	ignored	ignored
	H-	btone	btone	ignored	ignored	ignored
	!H-	btone	btone	ignored	ignored	ignored
	no label	none	none	ignored	ignored	ignored
	#Classes	2	2	5	2	3
State of the Art	Reference	[15]	[4]	[3]	[8]	[6]
	Level	word	syllable	word	word	syllable
	#Words/Syllables	24,955	44,390	29,578	29,800	14,599
	#Speakers	4	6	6	6	1
	Accuracy	**93.1%**	**91.61%**	**72.91%**	**67.7%**	**66.9%**
This Work	Level	word		word	word	word
	#Words	29,902		29,902	29,902	29,902
	#Speakers	6		6	6	6
	Accuracy	**89.0%**		**80.1%**	**83.1%**	**80.6%**

University of Illinois at Urbana-Champaign has shown the efficiency of neural networks and GMMs [23]–[25] on the recognition of prosodic events. In [26], the authors apply Bayesian networks to simultaneously recognize the words and the prosodic tags. This combined strategy provides lower error rates than a standard speech recognizer.

Concerning other languages, there are studies for Japanese [27] and Korean [28] to predict J-ToBI and K-ToBI labels, respectively. Both systems were developed in the framework of text-to-speech applications, and they solve the prediction of the ToBI labels from the text and from the acoustic signal.

From this analysis, we conclude that more work has been performed in the field of ToBI event detection than in classification. Some of the referenced authors explicitly mention the difficulty in the classification task, and they focus on the easier task of detection. In this work, we present an alternative strategy for the prosodic event classification task, as described in the following subsection.

B. Experimental strategy and overview

The Boston University Radio News Corpus (BURNC) [12] is used in this paper. The quality of this corpus has been already assessed in several studies about prosody modeling and, in particular, in the ToBI event detection/classification problem. The use of a common corpus with respect to other state-of-the-art approaches is a must for contrasting results. Section II is devoted to describing the peculiarities of the BURNC corpus.

Different parameterization techniques will be combined. We extended the basic set of raw prosodic features to sophisti-

TABLE III
ACCURACY OF THE BREAK INDEX CLASSIFICATION FOR DIFFERENT MAPPINGS OF THE TOBI LABELS: A COMPARISON BETWEEN THE STATE OF THE ART AND OUR APPROACH. BOSTON UNIVERSITY RADIO NEWS CORPUS HAS BEEN USED IN ALL CASES.

		Detection	Classification
Mapping	0	NB	0
	1,1-,1p	NB	1
	2,2-,2p	NB	2
	3,3.,3p	B	3
	4,4-	B	4
	#Classes	2	5
State of the Art	Reference	[15]	[18]
	Level	word	word
	#Words	24,955	8,568
	#Speakers	4	1
	Accuracy	**84.0%**	**70.4%**
This Work	Level	word	word
	#Words	28,723	28,723
	#Speakers	6	6
	Accuracy	**88.1%**	**74.6%**

TABLE IV
THE NUMBER OF TOBI PROSODIC EVENTS OF THE BOSTON UNIVERSITY RADIO NEWS CORPUS USED IN EXPERIMENTS.

Pitch Accents	
H*	7,587
L+H*	2,383
!H*	2,144
H+!H*	586
L+!H*	638
L*	517
L*+H	44
none	13,868

Boundary Tones	
L-L%	3,240
!H-L%	20
H-L%	187
L-H%	2,203
H-H%	36
L-	896
H-	955
!H-	821
none	21,544

Break Indices	
0	724
1,1-,1p	17,475
2,2-,2p	2,384
3,3-,3p	2,628
4,4-	5,512

cated approximations such as the Bézier parameters (already presented in [9]). We further extended the features, with the inclusion of Tilt parameters [11], as detailed in Subsection II-A.

All of the BURNC speakers were processed. This decision implies the use of normalization techniques that will be explained in Subsection II-B. The techniques used to cope with the imbalanced input problem are also explained in this section.

Some of the classes are easier to identify than others, as shown in previous work [9]. Furthermore, not all of the classifiers have the same behavior when they attempt to discriminate different classes and, in some cases, their outputs are complementary. In this paper, we take advantage of these arguments to defend an alternative classification strategy that is presented in Section III. Subsection III-A describes how to divide the 1 to n classification task into a set of subtasks in which the classification is binary. The combination of binary decisions is defended as a practical solution for improving classification results. The classification of prosodic events must account for the context in which the events are located. In this paper, we analyze the features of the adjacent words in the classification of the tones of a given word, and we use the Viterbi algorithm to improve the classification according to n-gram models that were previously trained (Subsection III-B). Subsection III-C shows the different fusion methods that were used to combine the outputs of the classifiers. Subsection III-D presents the basic classifiers to be used.

The results presented in Section IV will analyze the improvements in the classification rates that were obtained by the introduction of each of the different parts of the labeling strategy. Aside from the high accuracy of results, the confusions of the classifiers are contrasted with the common confusions that were observed in inter-transcriber tests. Finally, the conditions of other state-of-the-art studies are repeated to obtain the results that are displayed in Tables I, II and III. These results are discussed in Subsection IV-E.

II. PROCESSING OF THE BOSTON RADIO NEWS CORPUS

The Boston University Radio News Corpus [12] includes labels that separate phonemes, syllables and words. Accents are marked with a ToBI label and a position. Table IV shows the Accent Tones, Boundary Tones and Break Indices that were considered in this paper. Other symbols such as *, *?, *X*?*, *%?*, *X%?*, *-?* and *-X?* were ignored. Other scarce tones, such as *L*+!H*, have also been discarded.

Inspired by previous studies [4], [18], we aligned the accent tones with respect to the prominent syllable and to the word that contains it (words with more than one label are discarded in this work). All of the experiments were performed by using the word as the reference unit. This decision assumes that the temporal evolution describing the different tones can overflow the syllable duration; this strategy is consistent with other studies that are state of the art.

All of the utterances in the corpus with ToBI labels from all of the speakers ($f1a$, $f2b$, $f3a$, $m1b$, $m2b$ and $m3b$) were used. The fact that different speakers can have different pitch and energy registers justifies the use of normalization techniques, described in Section II-B.

Features that were similar to the features that were used in other experiments reported in the bibliography were used [4]:

- Frequency features: within-word F0 range ($f0\_range$), difference between maximum and average within-word F0 ($f0\_maxavg\_diff$), difference between average and minimum within-word F0 ($f0\_minavg\_diff$), difference between within-word F0 average and utterance average F0 ($f0\_avgutt\_diff$).
- Energy features: within-word energy range (e_range), difference between maximum and average within-word energy (e_maxavg_diff), difference between average and minimum within-word energy (e_minavg_diff).
- Maximum normalized vowel nucleus duration from all of the vowels of the word ($vowel\_duration$). Normalization is performed for each vowel type.
- Pause duration after a word ($pause$). Used only for boundary tones and break indices.

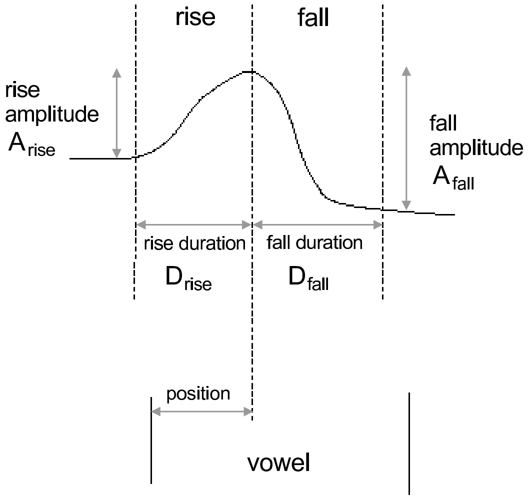

Fig. 1. Parameters of the RFC model used in the Tilt representation of an F0 contour. More details can be found in [30].

- Pseudo-grammatical information, part of speech (POS). We used the POS tags that come with the BURNC corpus, which were automatically obtained and were hand-corrected. The part-of-speech tags are the same as the Penn Treebank [29]. Some classifiers cannot handle qualitative features such as the POS feature; thus, quantization is necessary and is described in Section III-D1.

We will refer to all of these features as *raw parameters* in the remainder of this paper.

A. Parametric F0 contours

The *raw parameters* listed in the previous section have been shown to be useful for finding contrast between different linguistic units, such as syllables or words [4], [15]. This inter-unit contrast permits us to identify the emphatic function (related to the Pitch Accent) of a given unit or the cohesion between units (related to the Boundary Tones or Breaks). Nevertheless, these *raw parameters* appear to be insufficient in ToBI tone classification tasks. The different accents or boundaries contrast with each other in the temporal evolution of the pitch contour along the unit of reference. The accurate description of this temporal evolution requires a more sophisticated representation. In this paper, we use two of these representations: Tilt [11] and Bézier stylization [31].

Tilt is probably the most widely applied technique for parameterizing the pitch contours. This technique has its origin in the need to represent the relevant movements of the pitch contours in text-to-speech applications [33]. Moreover, it is the supporting technique of Festival[1], probably the most popular public domain text-to-speech system. The Tilt parameters are obtained by a combination of RFC parameters, with a numerical approximation of the prosodic events. Fig. 1 describes the parameters that are linked to a syllable in which a relevant prosodic event takes place, typically a Pitch Accent or a Boundary Tone. In this work, we use the 5 Tilt parameters: $F0\ height$, $position$, $amplitude$, $duration$ and $tilt$. Tilt has been explicitly used in some of the studies that were presented in the state-of-the-art section (Section I-A), such as [13] and [8].

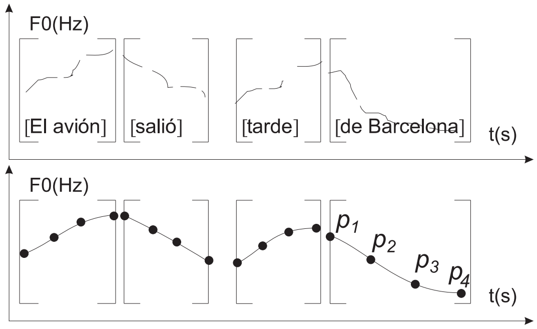

Fig. 2. An example of the Bézier function fitting stylization (from [32]). A set of quantitative parameters (p_1, p_2, p_3, p_4) represent the temporal evolution of the F0 contour in a given intonation unit (e.g., stress group, word, and syllable)

There exist alternatives to this representation; one of the alternatives was defended by our group in [32]. It is based on the approximation of the pitch contours with Bézier functions. The minimum square fitting approximation technique is used to represent the shape of the F0 contour along a given reference unit (Fig. 2 illustrates the process). 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 that have recently arisen, such as quantified contour modeling [3]. Both of these proposals have the advantage of allowing an increase in the number of parameters, in terms of the required accuracy.

B. Normalization and oversampling

Despite the fact that the input features are relative magnitudes, in [34] we showed that there are significant differences that affect the values of different speakers. For example, the feature $f0\_minavg\_diff$, which appears to be the most discriminative feature among the set of raw parameters for the identification of the accent, has a mean and standard deviation of 12.3 Hz and 13.8 Hz for the speaker $m1b$ and 30.9 Hz and 26.4 Hz for the speaker $f2b$. This situation also affects the remaining speakers and the input features. Under these conditions, a comparison of events that correspond to different speakers is risky and could potentially lead to confusing situations.

To minimize this effect, normalization techniques have to be applied. Several classical alternatives (Z-Norm, Min-Max and Euclidean 1-norm) were evaluated using neural network classifiers with raw, Bézier and Tilt features without context. Since Z-Norm normalization across the same speaker showed the best results, it was used thereafter.

Table IV shows that the number of samples per class is clearly imbalanced. This fact is not dependent on the corpus but instead results from the nature of the phenomenon itself.

[1] http://www.cstr.ed.ac.uk/projects/festival/

An English speaker naturally uses some prosodic events more frequently than others, which is very difficult to balance even with the preparation of an artificial ad-hoc corpus.

In [9], the negative impact of imbalanced data on the final result was shown, related to the fact that the classifiers tend to specialize themselves in the recognition of the most populated classes. In order to reduce this negative impact, several experiments applying undersampling, oversampling and ensemble sampling techniques to pitch accent type classification have also been reported in [13].

The approaches that are proposed for addressing imbalanced data can be divided into internal and external approaches, i.e., at the algorithmic or data level, respectively [35]. With the first approach, new algorithms or modifications of existing algorithms are proposed. In the second approach, the data sets are re-sampled, *over-sampling* the minority class or *under-sampling* the majority class. Both of these options can be accomplished randomly or directed, where examples to be generated or eliminated are respectively informed. We are interested in general solutions; thus, only external solutions have been evaluated. We performed several experiments and decided to use oversampling because it provided the best overall results for our classification system.

III. Classification procedure

As we already mentioned in the introduction, we showed in previous work [9] that the multiclass classification problem is a difficult task for a single classifier and that high accuracy results can be obtained based on the automatic distinction of the prosodic labels in pairs. Furthermore, different types of classifiers appear to behave differently, depending on the prosodic label that is to be identified. These are the main arguments that lead us to propose the strategy of classification that we formalize in this section. The next section shows its efficiency.

A. Fusion of pairwise coupled classifiers

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 [36], [37]. According to this approach, let us refer by $\hat{P}(l|x, \lambda^k_{l,m})$ to an estimation of the probability $P(y = l|x, \quad y = l \vee m)$, where l and m are two different prosodic labels; x is the input of the classifier (in our case, the prosodic features described in the previous section); y is the class label; and $\lambda^k_{l,m}$ is a pairwise classifier of type k (in our case, a decision tree or a neural network) that is trained to separate classes l and m.

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

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

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

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

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

The system proposed can be seen graphically in Fig. 3. There are as many classifiers as there are combinations of pairs of C classes: $\frac{C \cdot (C-1)}{2}$. Each classifier, $\lambda^k_{l,m}$, provides the confidence scores $\hat{P}(l|x, \lambda^k_{l,m})$ and $\hat{P}(m|x, \lambda^k_{l,m})$, which can be used as an estimation of the posterior probabilities (in the rest of the paper, we will refer to them as posterior probability estimates). The results of the classifiers are fused, as described in (1) and (2). Finally, the classification rule selects the label l^* so that the following occurs:

$$l^* = \operatorname*{argmax}_l \hat{P}(l|x) \tag{3}$$

As an alternative to this classification rule, we introduce a language model dependence that will be explained in the next subsection. Section III-C details the fusion operators $\bigotimes$, and Section III-D presents each of the K classifiers that are used in this work.

B. Recognition of label sequences

The ToBI standard of intonation [1] describes intonational contours as a sequence of tones. As a consequence, the labeling of a given word is highly dependent on the context in which the word has been uttered. This fact implies that the acoustic prosodic features that are observed in the surrounding words must be considered in addition to the labels that are assigned to the words of the context.

Concerning the impact of the acoustic realization of the context, the experiments that are reported in [3], [13], [16] showed an improvement in the results when the context information was included. To incorporate this contextual effect in our work, every input x must include prosodic features that are extracted from the corresponding word and from the context. Thus, each utterance of the corpus is composed of a sequence of words $w_1, w_2, \ldots, w_N$ from which we obtain a sequence of feature vectors $x_1, x_2, \ldots, x_N$. Each feature vector x_i contains information that is extracted from the corresponding word w_i (the raw parameters and also the Tilt and Bézier parameters, as described in section II-A) and from adjacent words. Feature selection methods [38] are applied to reduce the cardinality of the vector x_i, as reported in section IV-B.

Concerning the impact of the sequence of labels that are assigned to the words in the context, experiments reported in [4], [18], [39] showed an improvement in results when a model of the sequence of labels was used. Given the sequence of feature vectors $\mathbf{X} = \{x_1, x_2, ..., x_N\}$, the objective is to find the best sequence of prosodic labels $\mathbf{L}^* = \{l_1, l_2, ..., l_N\}$, which makes a maximum of the probability $P(\mathbf{L}|\mathbf{X})$:

$$\mathbf{L}^* = \arg\max_L P(\mathbf{L}|\mathbf{X}) = \arg\max_L P(\mathbf{L})P(\mathbf{X}|\mathbf{L}) \tag{4}$$

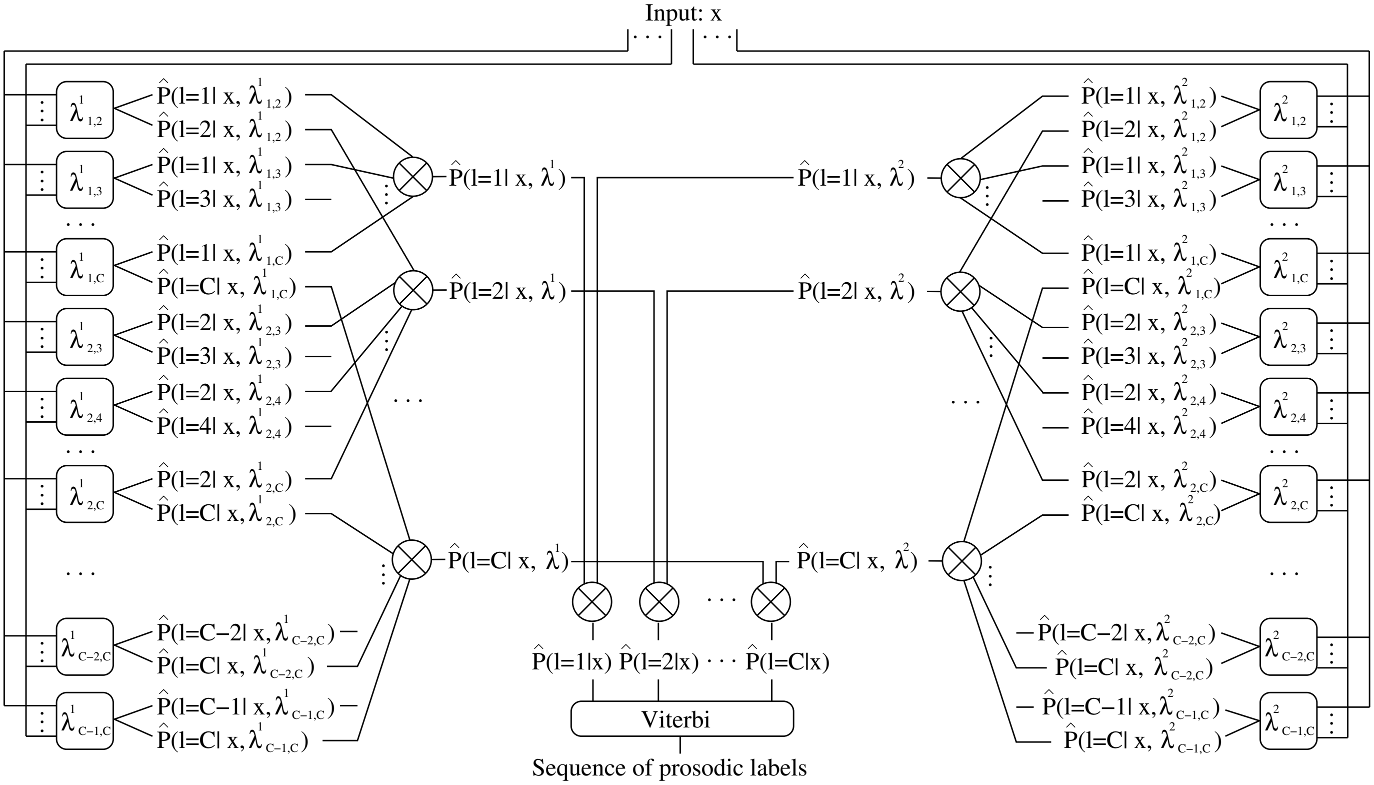
Input: x
$\lambda^1_{1,2}$
$\hat{P}(l=1| x, \lambda^1_{1,2})$
$\hat{P}(l=2| x, \lambda^1_{1,2})$
$\lambda^1_{1,3}$
$\hat{P}(l=1| x, \lambda^1_{1,3})$
$\hat{P}(l=3| x, \lambda^1_{1,3})$
$\lambda^1_{1,C}$
$\hat{P}(l=1| x, \lambda^1_{1,C})$
$\hat{P}(l=C| x, \lambda^1_{1,C})$
$\lambda^1_{2,3}$
$\hat{P}(l=2| x, \lambda^1_{2,3})$
$\hat{P}(l=3| x, \lambda^1_{2,3})$
$\lambda^1_{2,4}$
$\hat{P}(l=2| x, \lambda^1_{2,4})$
$\hat{P}(l=4| x, \lambda^1_{2,4})$
$\lambda^1_{2,C}$
$\hat{P}(l=2| x, \lambda^1_{2,C})$
$\hat{P}(l=C| x, \lambda^1_{2,C})$
$\lambda^1_{C-2,C}$
$\hat{P}(l=C-2| x, \lambda^1_{C-2,C})$
$\hat{P}(l=C| x, \lambda^1_{C-2,C})$
$\lambda^1_{C-1,C}$
$\hat{P}(l=C-1| x, \lambda^1_{C-1,C})$
$\hat{P}(l=C| x, \lambda^1_{C-1,C})$
$\hat{P}(l=1| x, \lambda^1)$
$\hat{P}(l=2| x, \lambda^1)$
$\hat{P}(l=C| x, \lambda^1)$
$\hat{P}(l=1| x, \lambda^2)$
$\hat{P}(l=2| x, \lambda^2)$
$\hat{P}(l=C| x, \lambda^2)$
$\hat{P}(l=1|x)$ $\hat{P}(l=2|x)$ · · · $\hat{P}(l=C|x)$
Viterbi
Sequence of prosodic labels
$\hat{P}(l=1| x, \lambda^2_{1,2})$
$\hat{P}(l=2| x, \lambda^2_{1,2})$
$\lambda^2_{1,2}$
$\hat{P}(l=1| x, \lambda^2_{1,3})$
$\hat{P}(l=3| x, \lambda^2_{1,3})$
$\lambda^2_{1,3}$
$\hat{P}(l=1| x, \lambda^2_{1,C})$
$\hat{P}(l=C| x, \lambda^2_{1,C})$
$\lambda^2_{1,C}$
$\hat{P}(l=2| x, \lambda^2_{2,3})$
$\hat{P}(l=3| x, \lambda^2_{2,3})$
$\lambda^2_{2,3}$
$\hat{P}(l=2| x, \lambda^2_{2,4})$
$\hat{P}(l=4| x, \lambda^2_{2,4})$
$\lambda^2_{2,4}$
$\hat{P}(l=2| x, \lambda^2_{2,C})$
$\hat{P}(l=C| x, \lambda^2_{2,C})$
$\lambda^2_{2,C}$
$\hat{P}(l=C-2| x, \lambda^2_{C-2,C})$
$\hat{P}(l=C| x, \lambda^2_{C-2,C})$
$\lambda^2_{C-2,C}$
$\hat{P}(l=C-1| x, \lambda^2_{C-1,C})$
$\hat{P}(l=C| x, \lambda^2_{C-1,C})$
$\lambda^2_{C-1,C}$

Fig. 3. Diagram of the classification procedure that is used in experiments.

The probability $P(\mathbf{L})$ is the probability of the sequence of labels $\mathbf{L}$. This probability is estimated, using the probabilistic language model, as the product of the probabilities of each label, assuming that the occurrence of a label is determined by the preceding labels:

$$\begin{aligned} P(\mathbf{L}) &= P(l_1, l_2, ..., l_N) \\ &= P(l_1)P(l_2|l_1)...P(l_N|l_1, l_2, ..., l_{N-1}) \\ &= \prod_{i=1}^{N} P(l_i|l_1, l_2, ..., l_{i-1}) \end{aligned} \tag{5}$$

where $P(l_i|l_1, l_2, ..., l_{i-1})$ is the probability of l_i, given that sequence $l_1, l_2, ..., l_{i-1}$ appeared previously. Using a trigram language model, the labels depend only on two previous labels; thus, (5) becomes the following:

$$P(\mathbf{L}) = P(l_1)P(l_2|l_1)\prod_{i=3}^{N} P(l_i|l_{i-2}, l_{i-1}) \tag{6}$$

The probability $P(\mathbf{X}|\mathbf{L})$ in (4) is the probability that the sequence of prosodic labels $\mathbf{L}$ produces an observation $\mathbf{X}$ and is calculated by the classifiers that are described in the previous section. From (2), the classifiers calculate the estimations $\hat{P}(l_i|x_i)$ for all of the possible prosodic labels that could be assigned to the word w_i.

Finally, (4) can be further developed, as described in [18] and [4]:

$$\begin{aligned} \mathbf{L}^* &= \arg\max_{L} P(x_1|l_1)P(l_1)P(x_2|l_2)P(l_2|l_1) \\ &\quad \cdot \prod_{i=3}^{N} P(x_i|l_i)P(l_i|l_{i-2}, l_{i-1}) \\ &= \arg\max_{L} \alpha(l_1|x_1)P(l_1)\alpha(l_2|x_2)P(l_2|l_1) \\ &\quad \cdot \prod_{i=3}^{N} \alpha(l_i|x_i)P(l_i|l_{i-2}, l_{i-1}) \end{aligned} \tag{7}$$

where

$$\alpha(l_i|x_i) = \frac{P(x_i|l_i)}{P(x_i)} = \frac{P(l_i|x_i)}{P(l_i)} \tag{8}$$

in our case, we use the estimations from the classifiers, as follows:

$$\alpha(l_i|x_i) = \frac{\hat{P}(l_i|x_i)}{P(l_i)} \tag{9}$$

The problem of finding the best sequence of prosodic labels is similar to the speech recognition problem. To search for the most likely prosodic label sequence, we built a graph that represents the state space [40]. Then, we applied the Viterbi algorithm [41] to obtain the best sequence of labels $\mathbf{L}^* = \{l_1, l_2, ..., l_N\}$:

Initialization
$V_1(i) = \pi_i b_i(x_1) \qquad 1 \leq i \leq M$

$B_1(i) = 0$

Recursion

$V_t(j) = \max_{1 \le i \le M}[V_{t-1}(i)a_{ij}]b_j(x_t) \qquad 2 \le t \le N, \quad 1 \le j \le M$

$B_t(j) = \arg\max_{1 \le i \le M}[V_{t-1}(i)a_{ij}] \qquad 2 \le t \le N, \quad 1 \le j \le M$

Termination

$s_N^* = \arg\max_{1 \le i \le M}[B_N(i)]$

Backtracking

$s_t^* = B_{t+1}(s_{t+1}^*) \qquad t = N-1, N-2, \ldots, 1$

where π_i is the initial state probability for state i, and a_{ij} is the transition probability from state i to state j. Both probabilities are obtained from the language model. The probability $b_i(x)$ is the probability that feature vector x occurs in state i and is calculated using the classifiers that are described in the previous section. M is the number of states. If C is the number of all possible values for a prosodic label l_i, then the state space has $M = C^2$ states when using a trigram language model. The best sequence of labels $\mathbf{L}^* = \{l_1, l_2, ..., l_N\}$ is obtained from the sequence of states $\mathbf{S}^* = \{s_1, s_2, ..., s_N\}$.

The SRILM toolkit was used to build trigram prosodic language models [42], with Katz backoff for smoothing. The training data was used to build these models.

C. Score Fusion

The general problem that is approached here is the so-called post-classifier fusion problem, in which the information is combined after the outputs of the classifiers have been obtained [43]. This integration of information can be divided into the following categories:

1) Fusion at the measurement level, using classifiers (for example, Neural Networks, SVM or Decision Trees) or using a combination approach (where the individual scores are combined by means of the sum or the product operation, to obtain a final score).
2) Fusion at the rank level, if the output of each matcher is a subset of the possible matches that are sorted in decreasing order of confidence.
3) Fusion at the abstract or decision level, using, for example, the majority voting technique or the AND or OR rules.

Here, the evidence (output) that is provided by each classifier is the information to be fused, with operator $\bigotimes$ in (1) and (2). Then, the second and third approaches from the previous list cannot be applied. From the options of the first category, the use of classifiers was discarded to simplify this stage. Then, we focused on the combination approach.

Several fusion methods [43] were tested in the preliminary experiments, to compare their performances. The combination techniques were evaluated as follows:

- Minimum value: $\hat{P}(l|x) = \min_k\{\hat{P}(l|x, \lambda^k)\}$.
- Maximum value: $\hat{P}(l|x) = \max_k\{\hat{P}(l|x, \lambda^k)\}$.
- Product: $\hat{P}(l|x) = \prod_{k=1}^{K} \hat{P}(l|x, \lambda^k)$.
- Sum: $\hat{P}(l|x) = \sum_{k=1}^{K} \hat{P}(l|x, \lambda^k)$.
- Sum using extreme values: $\hat{P}(l|x) = \min_k\{\hat{P}(l|x, \lambda^k)\} + \max_k\{\hat{P}(l|x, \lambda^k)\}$, which showed good performance in a previous study [44].
- Weighted sum, which can include class-independent weighting $\hat{P}(l|x) = \sum_{k=1}^{K} c_k \times \hat{P}(l|x, \lambda^k)$ and class-dependent weighting $\hat{P}(l|x) = \sum_{k=1}^{K} b_{kl} \times \hat{P}(l|x, \lambda^k)$. There are several approaches to obtaining the coefficients [43], [45]. Here, the fusion/calibration toolkit FoCal multi-class by Niko Brummer[2] has been used to obtain the coefficients (class-dependent). This toolkit includes discriminative logistic regression as well as generative Gaussian back ends, with PPCA, factor-analysis and HLDA covariance regularization [46].

The results that were achieved in preliminary experiments showed that the performance of the product of probabilities was similar to or only slightly worse than the best results that were obtained with the other fusion methods (usually achieved with the Sum). These results, together with the probabilistic interpretation of the product, which allows the application of Viterbi, was the reason for choosing the product of the probabilities as the fusion method.

D. The classifiers

We use two different types of classifiers in this work: decision trees (DT) and neural networks. Decision trees have been broadly used in the representation of the correspondence between the shape and the function of intonation because of their capability to combine qualitative and quantitative variables in a common framework. Specifically, for predicting ToBI accents, decision trees were used in the pioneering works of Ostendorf, such as in [6], [18]. Neural networks have also been used with this aim. In [47], an example is provided for the application of neural networks to predicting intonation in text-to-speech applications, with an extensive review of the use of neural networks in modeling intonation.

Of course, many other classification techniques can be found in the state of the art [48], but our intention was not to be exhaustive in the use of different types of classifiers. The reason for selecting these two types of classifiers is that, in previous works [9], we observed that classifiers behave differently on the discrimination of prosodic labels. Indeed, the decision tree seemed to specialize in the most populated classes, while the neural network balanced better the accuracy of the output for all of the classes (this result is observed again in this paper, in Section IV-A). The fusion of the scores of both of the classifiers improves the result using the fusion strategy that is shown in (2); therefore, $\hat{P}(l|x, \lambda_l^1)$ are the estimations that are obtained using the neural network, and $\hat{P}(l|x, \lambda_l^2)$ are the estimations that are obtained using the decision tree. Next, we present the implementation details of both of the classifiers.

1) The Multilayer Perceptron: A multilayer perceptron (MLP) is used, which is trained by means of the standard Error Backpropagation learning algorithm. Non-linear sigmoid units

[2] http://sites.google.com/site/nikobrummer/focalmulticlass

TABLE V

ACCURACY OF THE PITCH ACCENT TONE CLASSIFICATION USING DECISION TREES AND RAW, BÉZIER AND TILT FEATURES. THE TABLE ON THE LEFT SHOWS THE ACCURACY USING A SINGLE DECISION TREE. THE TABLE ON THE RIGHT SHOWS THE ACCURACY OF THE PAIRWISE CLASSIFIERS, WHERE POSITION l, m OF THE TABLE REPRESENTS THE TOTAL SUCCESS RATE OF THE CLASSIFIER $\lambda^2_{l,m}$.

Decision Tree	
H*	50.1%
L+H*	27.8%
!H*	20.8%
H+!H*	11.8%
L+!H*	8.0%
L*	17.4%
L*+H	0.0%
none	79.6%

	Pairwise classifiers							
	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*								
L+H*	67.4%							
!H*	71.2%	69.8%						
H+!H*	89.0%	83.2%	74.0%					
L+!H*	85.5%	69.6%	69.8%	75.4%				
L*	92.9%	89.1%	83.1%	67.7%	83.3%			
L*+H	98.7%	96.3%	96.2%	89.8%	90.5%	89.8%		
none	84.9%	92.3%	88.6%	93.3%	95.7%	94.2%	99.3%	

are used in the hidden and output layers because they showed better performance than $tanh()$ units in our experiments. Several network configurations were tested, to define the final MLP configuration:

- Single hidden layer and a total of 100 training epochs.
- Following results in Gori [49], more hidden units than inputs were used, to achieve the goal that the separation surfaces between classes in the pattern space can be closed.
- As many units as classes are used in the output layer: one per each class to classify.
- To train the MLP, unsaturated desired outputs [50] were tested. The chosen outputs, however, were 1.0 for the output that corresponded to the training vector class and 0.0 for the remainder because a better performance was achieved.

The MLP cannot address non-numerical inputs; in those cases, the POS feature must be numerically encoded. The following alternatives were tested:

- **POS Quantization (PQ)**. The $[0, 1]$ interval[3] is divided into p parts, with p the number of POS feature values, for which each of these POS values is assigned the lower numerical value of each of the p subintervals.
- **POS Codification (PC)**. Each of the POS feature values is codified in binary, using b bits ($p \leq 2^b$).
- **Input per POS (IP)**. An input is assigned to each POS value. Then, each POS feature is replaced by p new features, which have values that are 0 except for the value that is assigned to the corresponding POS value, which is 1.

The best results were obtained with the PC technique; thus, we transformed the POS tags into quantitative characteristics by using the codification of the 33 values, using 6 bits.

In [51], it was demonstrated that, given a discriminative neural network that was trained to distinguish between n classes (each output is assigned to one class, $O(l, x)$), no matter what the details of the structure of the neural network are, the global optimum is obtained if the outputs of the neural network are exactly the a posteriori probability $P(l|x)$, with x the input vector. In practice, it is impossible to know whether and how closely this optimum can be obtained because this task depends on the structure of the neural network and the number of training examples. In general, the neural network does not have a sufficient number of free parameters or degrees of freedom to produce exactly the class probabilities. However, the network has been trained so that the so-called *stochastic constraints* on the network outputs have been imposed, which implies that these network outputs can be used directly as class probabilities [51]. Then, although usually the probabilities provided by the neural network are different from the true class probabilities, $\hat{P}(l|x) = O(l, x)$ can be used as an estimation. Going back to (1), the MLPs are used as binary classifiers, so that $\hat{P}(l|x, \lambda^1_{l,m}) = O(l, x, \lambda^1_{l,m})$ and $\hat{P}(m|x, \lambda^1_{l,m}) = O(m, x, \lambda^1_{l,m})$ for every pair of labels $l, m = 1..C; l \neq m$.

2) C4.5 Decision Tree: The Weka machine learning toolkit [52] was used to build C4.5 decision trees (J48 in Weka). Different values for the confidence threshold for pruning have been tested, although the best results are obtained with the default value (0.25). The minimum number of instances per leaf is also set to the default value (2). This classifier was trained with qualitative POS features and unnormalized data. We decided to use unnormalized data because using decision trees the results obtained with unnormalized data were similar to those obtained with normalized data.

To obtain better class probability estimates, we turned off pruning, turned off *collapsing* and calculated class probabilities with the Laplace correction, as described in [53]. The decision trees will be trained as binary classifiers, to obtain the estimation of the probabilities $\hat{P}(l|x, \lambda^2_{l,m})$ and $\hat{P}(m|x, \lambda^2_{l,m}); l, m = 1..C; l \neq m$, to be fused, as explained in the previous section.

IV. EXPERIMENTAL RESULTS

In all the experiments we applied 10-fold cross-validation. The folds were created by randomly assigning paragraphs to folds. We decided to use paragraphs to divide the data in order to be able to recognize label sequences using the Viterbi algorithm, as described in section III-B.

Table V contrasts the classification results of the multiclass scenario compared to the results that were obtained by the pairwise classifiers. The rates that were obtained by pairwise classifiers were higher than 80% in more than 70% of the pairs of classes. These rates decrease in the multiclass scenario,

[3] The MLP inputs must be in the range $[0, 1]$ for a better network performance.

TABLE VI
The accuracy of the pitch accent classification using different combinations of features and different classifiers (raw: raw parameters; bez: Bézier parameters; tilt: tilt parameters).

	Decision Tree				**Neural Network**			
	raw	raw+bez	raw+tilt	raw+bez+tilt	raw	raw+bez	raw+tilt	raw+bez+tilt
H*	46.0%	47.1%	49.3%	50.1%	23.1%	22.9%	22.8%	24.1%
L+H*	21.3%	24.2%	26.9%	27.8%	34.8%	42.1%	41.5%	45.2%
!H*	17.8%	20.5%	20.8%	20.8%	20.8%	25.5%	29.3%	31.3%
H+!H*	11.3%	12.8%	11.4%	11.8%	33.4%	42.2%	42.3%	43.5%
L+!H*	5.3%	7.7%	5.8%	8.0%	32.8%	31.5%	29.2%	34.8%
L*	10.1%	15.9%	12.2%	17.4%	49.9%	54.2%	47.4%	55.9%
L*+H	0.0%	2.3%	0.0%	0.0%	2.3%	2.3%	2.3%	2.3%
none	75.6%	76.9%	79.4%	79.6%	65.8%	64.9%	68.5%	67.9%
Total	54.1%	55.7%	57.7%	58.2%	46.1%	46.9%	48.7%	49.5%

TABLE VII
The accuracy of the boundary tone classification using different combinations of features and different classifiers (raw: raw parameters; bez: Bézier parameters; tilt: tilt parameters).

	Decision Tree				**Neural Network**			
	raw	raw+bez	raw+tilt	raw+bez+tilt	raw	raw+bez	raw+tilt	raw+bez+tilt
L-L%	63.9%	68.7%	65.3%	68.5%	63.4%	67.8%	66.2%	68.8%
!H-L%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	15.0%
H-L%	7.0%	7.0%	7.0%	8.6%	18.2%	24.6%	21.4%	21.4%
L-H%	34.1%	44.6%	36.2%	45.5%	30.0%	48.4%	33.5%	48.1%
H-H%	0.0%	5.6%	8.3%	8.3%	0.0%	27.8%	19.4%	11.1%
L-	11.9%	14.6%	13.8%	15.4%	45.3%	52.5%	46.5%	50.0%
H-	12.4%	15.1%	16.8%	19.2%	34.8%	48.2%	44.5%	52.6%
!H-	11.7%	11.2%	12.7%	12.1%	22.7%	35.1%	35.3%	40.6%
none	84.9%	85.3%	85.0%	85.4%	57.1%	55.3%	57.6%	58.1%
Total	71.7%	73.4%	72.3%	73.8%	53.4%	55.0%	55.1%	57.3%

TABLE VIII
The accuracy of the break index classification using different combinations of features and different classifiers (raw: raw parameters; bez: Bézier parameters; tilt: tilt parameters).

	Decision Tree				**Neural Network**			
	raw	raw+bez	raw+tilt	raw+bez+tilt	raw	raw+bez	raw+tilt	raw+bez+tilt
0	8.3%	9.8%	9.5%	9.3%	57.2%	52.3%	53.0%	50.0%
1,1-,1p	72.2%	73.0%	73.5%	73.9%	29.5%	34.2%	34.4%	36.7%
2,2-,2p	16.3%	16.8%	18.9%	18.1%	32.9%	37.4%	36.6%	40.4%
3,3-,3p	21.8%	21.8%	24.0%	24.4%	54.4%	53.7%	52.6%	51.1%
4,4-	72.0%	72.9%	72.4%	72.5%	70.4%	72.3%	72.1%	73.4%
Total	61.3%	62.0%	62.6%	62.8%	40.6%	44.0%	44.0%	45.7%

where more than 75% of the classes have a rate that is lower than 50%. This result demonstrates the potential of the pairwise-coupled multiclass classification approach, and in this section, we experimentally show this fact.

First, Section IV-A shows the impact of the use of the Bézier and Tilt features for the recognition of different classes and the different behavior of the two types of classifiers that were used. Next, the importance of the context features is presented in Section IV-B. Then, we present in section IV-C the results of the pairwise fusion approach. Next, the errors of the classification are analyzed in section IV-D. Finally, our results are compared with the state-of-the-art in section IV-E.

A. Impact of different input features and classifiers

Tables VI, VII and VIII show the importance of using input features that characterize properly the evolution of the F0 contour in the words. The use of Bézier features improves the total classification rate in all of the cases. The most significant increase occurs in breaks using neural networks, where the rate goes from 40.6% to 44.0%. The use of Tilt parameters results in accuracy rates that are comparable to Bézier parameters for the boundaries and breaks. For the pitch accents, the Tilt parameters appear to represent more accurately the F0 contours.

The classification rate is highly dependent on the class. In the case of pitch accents and decision trees, it scores 50.1% for the accent *H** while accent *L** receives a 17.4%. More important is that the relative improvements that are obtained with the inclusion of the Bézier or Tilt features are also dependent on the class. Thus, the Bézier parameters appear to be more efficient in discriminating the *L** accent (from 10.1% to 15.9%) at the time that the Tilt is better at identifying other accents, such as *L+H** (from 21.3% to 26.9%). The last column of the tables combines both the Bézier and Tilt parameters and shows an improvement of the results in all of the cases.

TABLE IX
THE TOTAL CLASSIFICATION ACCURACY OF THE PITCH ACCENTS (ACCENT), THE BOUNDARY TONES (BTONE) AND THE BREAK INDICES (BREAK) USING THE FEATURE SETS WITH DIFFERENT CONTEXT FEATURES AND DIFFERENT CLASSIFIERS.

	Decision Tree			**Neural Network**		
Context	accent	btone	break	accent	btone	break
without context	58.2%	73.8%	62.8%	49.5%	57.3%	45.7%
1 previous word	60.4%	73.0%	62.8%	55.0%	57.8%	47.2%
2 previous words	60.1%	73.3%	62.7%	59.7%	63.8%	53.0%
1 following word	58.5%	74.5%	64.8%	53.1%	65.3%	56.9%
2 following words	57.9%	74.5%	64.8%	54.5%	69.1%	58.7%
1 prev. w. & 1 foll. w.	59.6%	74.5%	64.7%	59.8%	69.0%	57.9%
2 prev. w. & 2 foll. w.	60.1%	74.3%	65.3%	60.9%	71.7%	59.6%

TABLE X
THE TOTAL CLASSIFICATION ACCURACY OF THE PITCH ACCENTS (ACCENT), BOUNDARY TONES (BTONE) AND BREAK INDICES (BREAK) FOR DIFFERENT EXPERIMENTS. (DT: DECISION TREE; NN: NEURAL NETWORK)

CLASSIFIER VARIANT		accent	btone	break
Baseline (chance)		49.9%	72.1%	60.8%
Multiclass, no context	DT	58.2%	73.8%	62.8%
	NN	49.5%	57.3%	45.7%
Multiclass	DT	60.1%	74.3%	65.3%
	NN	60.9%	71.7%	59.6%
Pairwise coupling	DT	67.5%	81.5%	69.2%
	NN	68.1%	82.1%	70.5%
DT&NN Fusion	DT+NN	70.4%	83.9%	73.5%
Label Sequence (Viterbi)	DT+NN	**70.8%**	**84.2%**	**74.6%**

Another conclusion is that both of the classifiers behave differently with respect to different classes. The decision tree obtains a higher global accuracy than the neural network: 58.2% vs. 49.5% for accents; 73.8% vs. 57.3% for boundaries; and 62.8% vs. 45.7% for breaks. The decision tree is more efficient because it specializes in the most populated classes: *none* and *H** for accents; *L-L%* and *none* for boundary tones; and *1* and *4* for breaks. For these classes, the accuracy of the decision tree is greater than the accuracy that is obtained by the neural network; however, for the remainder of the classes, the neural networks obtain better results.

The different behaviors of the classifiers with respect to the different input features and also with respect to the different classes are the main evidence that justify the classifier fusion strategy that is detailed in the previous section; the results will be reported in Section IV-C.

B. Impact of the context

In this section, we analyze whether the use of context information can improve the classification results. We performed experiments to evaluate the impact of including context features and tried different configurations, in which the information from the two previous and two following words was used.

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 attribute selection with the Correlation-based Feature Selection (CFS) algorithm [38]. This method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. This approach helps to reduce the number of features to be used for the context.

Without the use of context, for each word, we use the 18 features in the pitch accent classification, as well as the 19 features in the boundary tone and break index classification. The CFS algorithm selected 8 features for the pitch accent classification: $f0\_minavg\_diff$, $f0\_avgutt\_diff$, e_range, POS and Tilt coefficients 2, 3, 4 and 5 ($position$, $amplitude$, $duration$ and $tilt$). The CFS algorithm selected 5 features for the boundary tone and break index classification: $vowel\_duration$, $pause$, POS, Bézier coefficient 4 and Tilt coefficient 4 ($duration$).

Table IX shows the results of using 6 different configurations of context: 1 previous word; 2 previous words; 1 following word; 2 following words; 1 previous word and 1 following word; and 2 previous words and 2 following words.

The inclusion of context features significantly improves the results. A similar improvement was also obtained in the experiments reported in [16], [54]. The impact of the context is different for each classifier: while the improvement in the decision tree is slight (from 62.8% to 65.3% in the best scenario), there is a large impact on the results of the neural network classification (from 57.3% to 71.7% for the boundary tones). The results also show that, for the pitch accent classification, the previous context causes more improvement than the following context. This result is consistent with the experiment described in [16]. The opposite situation occurs for the boundary tone and the break index classification, where the following context outperforms the previous context.

The best configuration is using 2 previous words and 2 following words as context in all of the cases; thus, the experiments in the following sections use this configuration.

C. Results of the final system

Table X shows the total classification accuracy that was obtained from different experiments. First, the *baseline* results are calculated based on chance: a recognizer that always assigns the majority class label (see Table IV for the distribution of events in the corpus). Second, a *multiclass* classification is performed, comparing the results that were obtained using a Decision Tree and a Neural Network and evaluating the impact of the context features. The following experiments evaluated the performance of the classification procedure that was proposed in this paper (see Fig. 3). First, we evaluated the improvement that was obtained by *pairwise coupling*, as described in (1). Next, we made the *fusion* of the Decision Tree and the Neural Network classifiers, as detailed in (2). Finally, we applied the Viterbi algorithm to find the best *label sequence*, as reported in section III-B.

The final results show that the classification procedure that is proposed in this paper outperforms the multiclass approach. There is a significant improvement in accuracy, going from 49.5% to 70.8% for the Pitch Accents, from 57.3% to 84.2% for the Boundary Tones and from 45.7% to 74.6% for the Break Indices.

The largest increase in the classification rate is obtained when the fusion strategies are used, especially the pairwise coupled classifiers (up to 7 percentage points in most cases). The improvement that is obtained by the trigram is also effective in discarding the odd sequences.

D. Errors of the classification

Table XI presents the most common errors of the final system, compared with a manual labeling process, and contrasts its consistency with respect to other subjective tests that are found in the state of the art. The confusion between each pair of labels is calculated as follows: for each pair of labels (l_1, l_2), we count the number of times that the system assigns the l_2 label when the l_1 label was expected and the number of times that the system assigns the l_1 label when the l_2 label was expected. Then, we divide that count by the total number of misclassified labels. A similar procedure is used for the manual labeling.

The most confused pair of Pitch Accents is *H** vs. *L+H**, which accounts for 27.26% of the total confusions. This pair of symbols is also problematic for manual ToBI transcribers, as has been observed in the inter-transcriber consistency test (25.64% of the inter-pair confusion in [14], in accordance with the high conceptual similarity index, 4/0, which is set in the same reference). The next most confused pair is *H** vs. *none*, with 22.71% of the confusions. This situation is problematic because it causes a wrong identification of an accent. Nevertheless, this result is consistent with the results in the manual transcription tests and with the unclear conceptual similarity index, which reflects that the manual transcribers assume that the symbols can be easy to confuse in some situations. For the rest of the Pitch Accent pairs, a high inter-transcriber confusion and a high similarity of the pair leads to more system confusion with that pair. The exception is the pair *L** vs. *none*, which is identified by the transcribers as a dissimilar pair (1/3 index), but it is on the list of commonly misclassified pairs. Nevertheless, it agrees with the inter-transcriber reliability result, as this pair of symbols is frequently confused in manual labeling experiments.

Concerning the Boundary Tones (central table of Table XI), the most difficult task appears to be the identification of the intermediate phrase boundaries (*H-*, *!H-* and *L-* vs. *none*). These situations are also difficult to identify by manual transcribers, as shown by the high conceptual similarity indices (4/0 and 3/1). Intermediate phrase boundaries are also frequently confused with their respective Final Boundary Tone counterpart (*L-* vs. *L-L%*, *H-* and *!H-* vs. *L-H%*), but to a smaller degree. The main weakness of our system is the distinction of the pairs *L-H%* vs. *none* and *L-L%* vs. *none*, that is, the identification of the final Boundary Tone. This situation is problematic because the symbols *L-H%*, *L-L%* and *none* are the most frequent ones (see Table IV), and manual transcribers find them easy to distinguish (0.18% and 2.20% rates).

In the case of the Break Indices, *1* appears in most of the pairs of often-confused Breaks (bottom sub table of Table XI). We do not have the conceptual similarity index for the Breaks. In spite of this fact, the pairs *1* vs. *2*, *0* vs. *1*, and *3* vs. *4* are frequently confused. Moreover, *0*, *1* and *2* on the one side, and *3* and *4* on the other side are frequently grouped in state-of-the-art studies (see Table III). Our weakest result is the confusion between *1* and *4* (10.51% of the cases), which is consistent with the results that are obtained for Boundary Tones, where the confusion of *none* with *L-L%* and *L-H%* has also been observed.

TABLE XI
Most common errors of the system, compared with the manual labeling process that is described in [14]. The *Automatic Classification Error* is the percentage of confusion of the system between the pair of labels (we only report high confusion pairs). The *Manual Labeling Disagreement* is the percentage of confusion of manual labelers between the pair of labels (derived from the table that is named *ALL labelers-POOLED* in [14]). The *Conceptual Similarity Index*: S/D, where S are the number of experts (out of four) that consider the pair of labels to be similar, and D is the number of experts that consider the pair of labels to be dissimilar (obtained from [14]).

Pitch Accent Tone				
		Automatic Classification Error	Manual Labeling Disagreement	Conceptual Similarity Index
H*	L+H*	27.26%	26.64%	4/0
H*	none	22.71%	13.15%	2/2
H*	!H*	13.17%	4.82%	4/0
!H*	none	9.75%	6.57%	3/1
L*	none	4.39%	4.03%	1/3
H*	L+!H*	3.51%	4.38%	4/0
H+!H*	none	3.22%	1.75%	2/2
!H*	L+!H*	2.88%	2.80%	3/1
H*	H+!H*	2.71%	1.67%	2/2

Boundary Tone				
		Automatic Classification Error	Manual Labeling Disagreement	Conceptual Similarity Index
H-	none	15.70%	19.96%	3/1
!H-	none	14.90%	8.97%	3/1
L-	none	13.32%	21.43%	4/0
L-H%	none	12.94%	0.18%	2/2
L-L%	L-H%	11.65%	11.17%	4/0
L-L%	none	8.67%	2.20%	1/3
L-L%	L-	4.60%	17.95%	4/0
L-H%	L-	3.42%	1.28%	3/1
L-H%	H-	3.97%	0.92%	0/4
L-H%	!H-	1.69%	0.00%	0/4

Break index				
		Automatic Classification Error	Manual Labeling Disagreement	Conceptual Similarity Index
1,1-,1p	2,2-,2p	31.35%	n/a	n/a
1,1-,1p	3,3-,3p	24.02%	n/a	n/a
3,3-,3p	4,4-	14.43%	n/a	n/a
1,1-,1p	4,4-	10.51%	n/a	n/a
0	1,1-,1p	9.96%	n/a	n/a

E. Final system compared with the state of the art

It is not easy to compare the results of our experiments with the results that are described in the state of the art because each experiment is performed with a different experimental setup. The main differences are the following: the ToBI labels that are used or ignored and the mapping between labels; the corpus/subcorpus; the number of speakers; and the word or syllable level.

In this section, we describe experiments that we made to compare our proposal with experiments that are found in the bibliography that used the BURNC corpus. Those experiments were described in Section I-A and are summarized in Tables I, II and III, where our results are also shown, using the same mappings of the ToBI labels used by other researchers.

Although detection is not the goal of our work, this task is included in the tables to show that our system obtains comparable results to the state of the art, with better results in some cases.

In pitch accent classification, using the same 5 labels as in [3], we achieved 69.1%, compared with 63.99% that they reported. With the same 4 labels of [8], we obtained 63.9%, compared with 56.4% that they reported. Using the 4 labels for the experiments described in [6], [16], [17], we obtained 80.0%, compared with their accuracy of 80.17%, 81.3% and 87.17%, respectively. However, their experiments were at the syllable level and used only one speaker; thus, a direct comparison is not possible.

In boundary tone classification, using the labels of the experiments that are shown in [3], we report 80.1% while their accuracy was 72.91%. With the labels of the work described in [8], we obtained 83.1%, compared with 67.7% that they reported. Finally, using the labels in [6], we achieved 80.6%, in comparison with 66.9% from their work.

For break indices classification, we reported 74.6% accuracy, and the experiments in [18] obtained 70.4%.

In summary, we achieved results comparable with the state of the art, and in the majority of the cases, our results outperformed the results that are found in the bibliography. An exception was in the pitch accent classification task, where other researchers reported results on speaker-dependent experiments that were better than ours.

V. Conclusions

A new and effective proposal to address the complex task of automatic classification of prosodic events is presented in this paper. The use of pairwise coupling for multi-label classification together with the fusion of complementary classifiers and an adequate feature selection has allowed us to improve the results of the baseline system.

The use of input features that reflect the evolution of the intonation contour in the words is shown to be important for characterizing the ToBI events. Thus, the inclusion of Tilt parameters and Bézier parameters improves the classification rates by 4.1 points for Pitch Accents (from 54.1% to 58.2%, Table VI). However, the main improvements were achieved with the use of pairwise coupled classifications in conjunction with the use of evidence from the context and with the fusion of different classifiers, increasing the classification rate by 12.6 percent points (from 58.2% to 70.8%, in Table X). The results of the final system are superior to the state of the art in most of the situations that were evaluated.

An analysis of the misclassified samples shows that the types of mistakes made by the system do not differ significantly from the common confusions observed in manual ToBI inter-transcriber tests. The use of ToBI labeling tools introduces an element of objectivity that can be very useful for accelerating the work of manual labeling by the transcribers. The percentage of approximately 80% success in the classification tasks guarantees a certain level of reliability, although in some situations (where it is difficult to label prosodic events), transcribers' intervention is still essential.

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 *International Conference on Spoken Language Processing (ICSLP)*, 1992, pp. 867–870.

[2] M. Beckman and G. Elam, "Guidelines for ToBI labelling," 1997. [Online]. Available: http://www.ling.ohio-state.edu/research/phonetics/E_ToBI

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

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

[5] J. Pitrelli, M. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *International Conference on Spoken Language Processing (ICSLP)*, 1994, pp. 123–126.

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

[7] A. K. Syrdal, J. Hirshberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody," *Speech Communication*, no. 33, pp. 135–151, 2001.

[8] S. Ananthakrishnan and S. Narayanan, "Fine-grained pitch accent and boundary tone labeling with parametric F0 features," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2008, pp. 4545–4549.

[9] C. González-Ferreras, C. Vivaracho, D. Escudero, and V. Cardeñoso, "On the Automatic ToBI Accent Type Identification from Data," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2010.

[10] A. Botinis, B. Granstrom, and B. Moebius, "Developments and paradigms in intonation research," *Speech Communication*, vol. 33, pp. 263–296, July 2001.

[11] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

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

[13] A. Rosenberg, "Automatic Detection and Classification of Prosodic Events," Ph.D. dissertation, University of Columbia, USA, 2009.

[14] 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.

[15] 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.

[16] 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.

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

[18] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, October 1994.

[19] K. Ross and M. Ostendorf, "A dynamical system model for generating fundamental frequency for speech synthesis," *IEEE Transactions on speech and audio processing*, vol. 7, no. 3, pp. 295–309, 1999.

[20] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2005.

[21] A. Rosenberg, “AuToBI – A Tool for Automatic ToBI Annotation,” in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2010.

[22] J. Hirschberg, “Pitch accent in context: Predicting intonational prominence from text,” *Artificial Intelligence*, vol. 63, no. 1-2, pp. 305–340, October 1995.

[23] Y. Ren, S. Kim, M. Hasegawa-Johnson, and J. Cole, “Speaker-independent automatic detection of pitch accent,” in *Proceedings of Speech Prosody*, 2004.

[24] K. Chen, M. Hasegawa-Johson, A. Cohen, and J. Cole, “A Maximum likelihood Prosody Recognizer,” in *Proceedings of Speech Prosody*, 2004.

[25] K. Chen, M. Hasegawa-Johson, and A. Cohen, “An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model,” in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2004, pp. 509–512.

[26] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, T. Yoon, and S. Chavarria, “Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus,” *Speech Communication*, no. 46, pp. 418–439, 2005.

[27] N. Campbell, “Autolabelling japanese tobi,” in *International Conference on Spoken Language Processing (ICSLP)*, 1996, pp. 2399–2402.

[28] J.-S. Lee, B. Kim, and G. G. Lee, “Automatic corpus-based tone and break-index prediction using k-tobi representation,” *ACM Transactions on Asian Language Information Processing*, vol. 1, pp. 207–224, September 2002.

[29] M. Marcus, M. Marcinkiewicz, and B. Santorini, “Building a large annotated corpus of English: The Penn Treebank,” *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[30] P. Taylor and A. Black, “The Rise/Fall/Connection model of intonation,” *Speech Communication*, vol. 15, pp. 169–186, 1995.

[31] D. Escudero, V. Cardeñoso, and A. Bonafonte, “Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish,” in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 481–484.

[32] D. Escudero and V. Cardeñoso, “Applying data mining techniques to corpus based prosodic modeling speech,” *Speech Communication*, vol. 49, pp. 213–229, 2007.

[33] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[34] D. Escudero, C. González, C. Vivaracho, V. Cardeñoso, and L. Aguilar, “Caracterización acústica del acento basada en corpus: un enfoque multilingüe inglés/español,” in *Congreso en Fonética Experimental*, Octubre 2011.

[35] C. Vivaracho-Pascual and A. Simon-Hurtado, “Improving ANN Performance for Imbalanced Data Sets by Means of the NTIL Technique,” in *IEEE International Joint Conference on Neural Networks*, 18-23 July 2010.

[36] T. Hastie and R. Tibshirani, “Classification by pairwise coupling,” *The annals of Statistics*, vol. 26, no. 2, pp. 451–471, April 1998.

[37] T.-F. Wu, C.-J. Lin, and R. C. Weng, “Probability estimates for multi-class classification by pairwise coupling,” *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, December 2004.

[38] M. A. Hall, “Correlation-based feature subset selection for machine learning,” Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1998.

[39] A. Rosenberg, “Symbolic and Direct Sequential Modeling of Prosody for Classification of Speaking-Style and Nativeness,” in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011.

[40] F. Jelinek, *Statistical Methods for Speech Recognition*. The MIT Press, 1998.

[41] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[42] A. Stolcke, “SRILM – An Extensible Language Modeling Toolkit,” in *International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 901–904.

[43] A. Jain, K. Nandakumar, and A. Ross, “Score normalization in multimodal biometric systems,” *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[44] C. E. Vivaracho, J. Ortega-Garcia, L. Alonso, and Q. I. Moro, “Extracting the most discriminant subset from a pool of candidates to optimize discriminant classifier training,” *Foundations of Intelligent Systems*, pp. 640–645, 2003.

[45] J. M. Pascual-Gaspar, M. Faundez-Zanuy, and C. Vivaracho, “Fast on-line signature recognition based on VQ with time modeling,” *Engineering Applications of Artificial Intelligence*, vol. 24, no. 2, pp. 368–377, 2011.

[46] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, “Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, september 2007.

[47] K. Sreenivasa Rao and B. Yegnanarayana, “Intonation modeling for Indian languages,” *Computer Speech & Language*, vol. 23, no. 2, pp. 240–256, 2009.

[48] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[49] M. Gori, “Are multilayer perceptrons adequate for pattern recognition and verification?” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1121–1132, November 1998.

[50] S. Lawrence, I. Burns, A. Back, A. Chung Tsoi, and C. L. Giles, “Neural networks classification and prior class probabilities,” *Lecture Notes in Computer Science State-of-the-Art Surveys*, pp. 299–314, 1998.

[51] H. Ney, “On the probabilistic interpretation of neural network classifiers and discriminative training criteria,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 107–119, 1995.

[52] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[53] F. Provost and P. Domingos, “Tree induction for probability-based ranking,” *Machine Learning*, vol. 52, no. 3, pp. 199–215, 2003.

[54] A. Rosenberg and J. Hirschberg, “Detecting Pitch Accent at the Word, Syllable and Vowel Level,” in *HLT/NAACL*, 2009.

César González-Ferreras received the B.Sc. and M.Sc. degrees in computer science in 1998 and 2000 respectively, both from University of Valladolid (Spain). He received the Ph.D. degree in computer science in 2009 from the University of Valladolid.

In 2001 he joined the Department of Computer Science at University of Valladolid, where he is currently a lecturer. His research interests include spoken language processing, spoken dialog systems and prosody recognition.

David Escudero-Mancebo received the B.A. degree in computer science in 1993; the M.Sc. degree in computer science in 1996; and the Ph.D. degree in information technologies in 2002 in the University of Valladolid, Spain. In 1994 he followed an intensification in telecommunications engineering (electronic systems) in the Polytechnic School of the University of Valladolid.

He is Associate Professor of computer science in the University of Valladolid, Spain. He is co-author of several publications in the field of computational prosody, both concerning modeling of prosody for text-to-speech systems and prosodic labeling of corpora.

Carlos Vivaracho-Pascual obtained the M.S. degree in Physics in 1989 and the PhD degree “cum laude” in 2004, both from the University of Valladolid (Spain). He has been a university lecturer in this University since 1996. He has published close to 40 technical papers (most of them in the main congresses of his research field and outstanding journals) and some book chapters. His research interests are focused on biometric (mainly speaker and signatures recognition) and pattern recognition. He is now involved in a major pedagogical reform in the Computer Science School of the University of Valladolid, with some papers and workshops organized in this field.

Valentín Cardeñoso-Payo (M'02) received the M.S. degree in physics in 1984 and the Ph.D in physics in 1988, both from the University of Valladolid, Valladolid, Spain. In 1988, he joined the Departamento de Informática, Universidad de Valladolid, where he currently holds an associate professor position in computer languages and information systems. He has been the Director of the Research Group on Multimodal Human Computer Interaction Systems since its creation. His current research interests include machine learning techniques applied to human language technologies, human computer interaction and biometric person recognition. He has been the advisor of ten Ph.D works in speech synthesis and recognition, on-line signature verification, voice based information retrieval and structured parallelism for high performance computing.