Estimating tonal prosodic discontinuities in Spanish using HMM

Alejandro Bassi [a], Nestor Becerra Yoma [b,*], Patricio Loncomilla [b]

[a] *Department of Computer Sciences, Universidad de Chile, Av. Tupper 2007, P.O. Box 412-3, Santiago, Chile*
[b] *Department of Electrical Engineering, Universidad de Chile, Av. Tupper 2007, P.O. Box 412-3, Santiago, Chile*

Abstract

The tonal prosodic discontinuity estimation in Spanish is exhaustively modelled using HMM. Due to the high morphological complexity in Spanish, a relatively coarse grammatical categorization is tested in two sorts of texts (sentences from newspapers and a theatre play). The estimation of the type of discontinuity (falling or rising tones) at the boundary of intonation groups is assessed. The HMM approach is tested with: (a) modelling the observation probability with monograms, bigrams and full-window probability; (b) state duration modelling; (c) discriminative analysis of intermediate and final observation vectors and (d) penalization scheme in Viterbi decoding. The optimal configurations led to reductions of 3% or 5% in error detection. The estimation of the observation probability with monograms and bigrams leads to worse results than the ordinary full-window probability, although they provide better generalization. Nevertheless, the performance of the monograms and bigrams approximation can be enhanced if applied in combination with state duration constraints.

1. Introduction

The intelligibility and naturalness of a text-to-speech (TTS) system depend to a great extent on its prosodic rendering. The main acoustic parameters to be modulated are the intensity, the duration and the tone or pitch. At the lexical level, prosody is easily integrated, for example to generate the stress patterns of isolated words. However, the prosody of higher linguistic units can not be reduced to a mere concatenation of lexical patterns. When parsing a sentence, a human hearer can recognize a succession of phrases or clauses relying on acoustic cues that are typically marked by pauses and tonal variations. A TTS system should be able to produce such cues for a better perceived phrasing, with a special regard to intonation, since it is one of the most salient aspects.

A major goal of a TTS prosodic component at the sentence level is to assign a melodic contour to the sentence, which requires that it is properly segmented in intonation phrases. For some applications, as dialogue system, where the enunciated text is controlled by an answering module, the segmentation and the contour of each phrase can be generated at the source, as an additional input to

* Corresponding author. Tel.: +56 2 678 4205; fax: +56 2 695 3881.
E-mail address: nbecerra@ing.uchile.cl (N.B. Yoma).

the prosodic component (Kochanski et al., 2003). In the case of a free-text reading application, the task requires first to automatically estimate, from the text only, the expected positions where a proficient reader would introduce distinctive segmenting cues. Additionally, a prediction of the specific intonation patterns at phrase boundaries is much helpful for the final TTS synthesis.

This paper deals with a corpus based data-driven estimation of the tonal discontinuities that occur at the boundaries of intonation phrases. Both the position and type of boundary are considered. A Spanish corpus was collected and annotated to train a HMM model that is used to estimate the sought variables. The technique is extendable to other languages and can be applied as a first processing step for unrestricted text applications.

The typical prosodic patterns of Spanish at the sentence level and the segmentation approaches that may be applied to estimate them are explained in Section 2. Section 3 is devoted to the specific HMM model used in this work. Several experiments using different configurations of the model and their results are presented in Section 4. Section 5 provides some concluding remarks.

2. Modeling prosodic discontinuities

The prosodic model followed in this paper to describe Spanish intonation at the sentence level is inspired by the general patterns proposed in (Navarro, 1944). According to Navarro (1944), sentences are usually divided in *intonation groups* or *melodic units* that do not overlap from sentence to sentence. The intonation group is the shortest possible speech segment with individual meaning and with a given melodic contour. Intonation groups can further be subdivided into *stress groups* that depend mainly on the lexical units of the utterance and on the focus. Intonation groups generally coincide with *breath groups*, defined as the sequence of connected speech between two pauses. This can be explained by assuming that intonation groups are undivided units, so breathing is constrained to take place at their boundaries. However, breathing itself is not a prosodic marker and is not reliable enough to be used for boundary detection purposes. Breathing is constrained by intonation groups but not completely determined by them. It must be noticed also that the length of pauses may certainly be affected by the occurrence of breathing, which mars their usefulness to classify boundary types.

According to the Autosegmental-Metrical, AM, approach, intonation can be described by using an abstract phonological representation level independently of its phonetic implementation details (Gussenhoven, 2002). An AM model has separate tiers for segments and tones. The segmental structure has several hierarchical levels, ranging from utterances and intonational phrases to syllables and phonemes. Tones make reference to segments: in Spanish, pitch accents are associated to syllables and boundary tones to intonational phrases. This article deals with the latter hierarchical level. There is some controversy whether Spanish has only one level of intonational phrasing or two (Beckman et al., 2002). Generally, it can be assumed that phrasing levels are marked by the intensity of the prosodic cues. However, such quantitative differences are difficult to model. In this paper a qualitative approach considering only a single phrasing level is adopted to address the segmentation problem. Fig. 1 shows the intonation contours of several sentences. As can be seen, some typical patterns are distinguishable. For example, declarative sentences end with a falling intonation and their intermediate intonation groups with a raising intonation.

The ordinary melodic contours of intonation groups in Spanish are characterized by: an initial rising until the first stressed syllable or the syllable following it; then, a more or less uniform intonation; and, depending on the type of intonation group, different final tonal movements that contain the main intonation information. In (Navarro, 1944) five such tonal movements were considered: falling, half-falling, level, half-rising and rising. However, from the practical point of view, classifying tonal movements in those five categories is a difficult task. Due to this fact, a more achievable approach was adopted to recognize only two types of final tonal movements: falling and rising. In (Fant, 1984; Quilis, 1993) a similar binary distinction was adopted. However, it can be argued that a more detailed specification of the phonetic description within intonation groups is required for a final TTS application in order to improve naturalness. Nevertheless the abstract binary description proposed here is useful to determine what type of boundary must be generated.

Pauses are often optional; hence they are unreliable segmenting cues. Consequently, rather than referring to breaks or pauses, as done by well known segmenting approaches, this paper proposes to base the analysis on the observed intonation

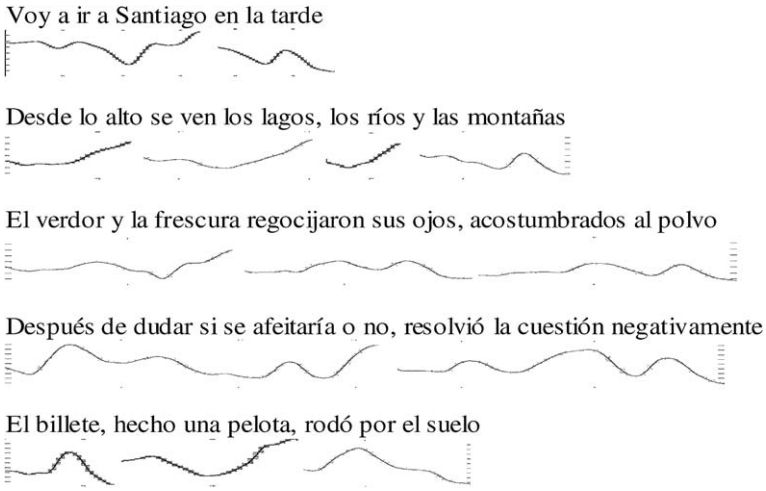

Fig. 1. Example of intonation contours in Spanish.

discontinuities. Since there are not systematic differences at the beginning of Spanish intonation groups, the discontinuities considered in this work are characterized according to the typical tonal movements of their final part that are reduced to two cases (see Fig. 1): first, a falling-tone (ToBI label, L%); or second, a rising-tone (ToBI label, H%).

Besides the perceptual relevance, tonal aspects of prosody at the sentence level are also important because they provide qualitative differences which allow improving the discrimination of phrase boundaries. Other prosodic markers such as the duration of vowels or pauses also contribute to segment the utterances. However, some studies show that it is difficult to use them to determine different boundary types (Garrido et al., 1995). As stated in (Garrido et al., 1995), "F0 resets seem to be the only prosodic cue that behaves in a different way according to the type of boundary" and "resets could be then related to major boundaries". Moreover, this intonation discontinuity is consistent with the perceived initial rising tonal movement observed in (Navarro, 1944) at the beginning of intonation groups.

The purpose of the current paper is to estimate intonation groups. The phonetic detail of how to generate them, including other prosodic markers besides tone, namely duration and amplitude, are out of the scope of this study. Moreover, a final TTS generation could be carried out using a rule based approach adapted to the target language or dialect. The underlying assumption is that the knowledge about the type of boundary should substantially help to determine all the other relevant parameters. The study presented in (Sosa, 1999) can contribute to define those parameters. Particularly, the use of L tones requires a detailed predictive model as described in (Prieto, 1998).

2.1. Automatic sentence segmentation approaches for TTS

In this paper, segmentation is defined as the labelling procedure of each possible place of the text (each word) to indicate the boundary condition. In contrast to previous published approaches, the model proposed in this paper does not deal with a two-class pause problem, *break* and *non-break*, but classifies word positions according to: label 0, non-discontinuity; label 1, falling-intonation ending; and, label 2, rising-intonation ending.

2.1.1. Knowledge-based approaches

The problem of automatically segmenting a text into intonation groups can be solved using a knowledge-based approach relying on rules generated by case dependent analysis (Anderson et al., 1984). A syntactic parser may be used to enhance this analysis (Atterer, 2002). The underlying assumption is that the prosodic segmentation reflects the syntactic structure of the text. However, the semantic and pragmatic aspects are very relevant but they can hardly be modelled. The rules try to describe the typical boundary conditions of the segments. Nevertheless, the detailed linguistic knowledge required by case-dependent approaches is difficult to

manage, which in turn may lead to a lack of robustness in real TTS systems. For instance, Spanish shows a highly complex morphology where verbs present many declinations. In contrast to knowledge-based approaches, trainable data-driven models are generally preferred in such applications. A comparative study for Spanish shows that data-driven techniques can provide good results (Agüero and Bonafonte, 2003).

2.1.2. The CART model

Classification and regression trees, CART (Breiman et al., 1984), have been applied successfully to this problem. A CART model is generated by starting with a single node to induce a full decision tree after a succession of appropriate incremental expansions of the leaves. Each node of the tree implements a simple partitioning rule using features such as part-of-speech tags and punctuation. This approach has been used for a local prediction of the position of pauses in Spanish (Hirschberg and Prieto, 1996). The main advantage of CART is the fact that it allows to handle symbolic features. Also, due to their hierarchical structure, in spite of a potentially very large combinatory, CART leads to compact solutions that can be interpreted as nested if-then-else rules.

2.1.3. Hidden Markov models (HMM)

HMM has widely been employed in speech recognition. HMM is a stochastic model where hidden sequences of states, resulting from a Markovian process, generate a sequence of observable outputs according to a probability distribution (Huang et al., 1990; Jelinek, 1998). In (Black and Taylor, 1997) the problem of estimating phrase boundaries was addressed with HMM. Two states were considered: break and non-break. Each observation is a vector that represents the syntactic context of a word. A typical choice for the observation vector is a narrow window of lexical items centred on the position of the word where the boundary state is evaluated. The components are chosen from a finite set of lexical tags, which implies a finite, although large, set of possible observations. The output probability is modelled with a discrete probability distribution. The sequence of breaks and non-breaks is estimated by using the Viterbi algorithm. The most important advantage of HMM over CART is the fact that the optimal assignment is evaluated over the whole utterance rather than locally (Black and Taylor, 1997).

2.2. Automatic part-of-speech (POS) tagging for TTS

Automatic sentence segmentation approaches for TTS require a proper linguistic tagging of the text in order to characterize the syntactic contexts. Each lexical unit of the text is labelled with the corresponding lexical category. Words equally labelled are expected to have the same grammatical behaviour. As a result of the automatic tagging process, each sentence is represented by the sequence of the lexical categories assigned to the words.

2.2.1. Punctuation marks and prosody

Punctuation marks do not belong to any syntactic class, but from a prosodic point of view they help to read a text and can be used as predictors of intonation group boundaries. However, in some cases the punctuation mark does not determine the intonation discontinuity univocally and further modelling is also required. The deterministic rules for intonation discontinuities extracted from the corpus employed here are summarized in Table 1. The problem of prosody discontinuity estimation within punctuation marks is always much more difficult and requires a more complex analysis. As can be seen in Table 1, some punctuation marks (e.g. "!", "?", etc.) univocally determine the pitch contour in the sentence. In contrast, the intonation discontinuities are not deterministically predicted when the sentences are ended by the following punctuation marks: ";", "," and ":".

In this paper, text is first divided into segments bounded by punctuation marks, which in turn define the type of intonation discontinuity in some cases according to deterministic rules. Then, a HMM based system is employed to break down the bounded segments into sequences of intonation groups and to evaluate the pitch curve within each group. Moreover, the system also estimates the intonation curve in those cases when the punctuation mark provides ambiguous information about the intonation curve as shown in Table 1. The HMM

Table 1
Punctuation marks and intonation discontinuities

Punctuation marks	Intonation discontinuity
"!" ; "." ; "". " ; ")." ; "-."	Falling-tone ending
"?"	Raising-tone ending
"," ; ";" ; ":" ; "…" , "", "	These punctuation marks do not define the intonation curve univocally

techniques employed in the results reported here are described in the following section.

3. HMM based prosodic discontinuity estimation

In contrast to Black and Taylor (1997), the HMM system employed here makes use of three states to model the intonation boundary conditions (Fig. 1): 0, non-discontinuity; 1, falling-tone ending; and 2, raising-tone ending. The HMM topology employed to model the prosody discontinuities mentioned above is shown in Fig. 2. According to Fig. 1, the states correspond to: state 0 (S_0), absence of intonation discontinuity; state 1 (S_1), falling-tone discontinuity; and, state 2 (S_2), raising-tone discontinuity. The HMM is defined by $\lambda = (A, B, \pi, \phi)$ where

$$A = \{a_{i,j} = \text{Prob}(\text{state } j \text{ at instant } t | \text{state } i \text{ at instant } t-1)\}$$
$$B = \{b_i(O_t) = \text{Prob}(O_t | \text{state } i)\}$$
$$\pi = \{\pi_i = \text{Prob}(\text{state } i \text{ at } t = 1)\}$$
$$\phi = \{\phi_i = \text{Prob}(\text{state } i \text{ at } t = T)\}$$

where T is the length of the observation sequence $O = [O_1, O_2, \ldots, O_t, \ldots, O_T]$ that corresponds to a text string or sentence after parameterization. As mentioned above, the proposed technique is tested with a Spanish database and the text strings or sentences correspond to the text between two consecutive punctuation marks as mentioned in Section 2.2.1. The parameterization procedure is described as follows:

(P1) Every word and punctuation mark in the text string is classified according to a lexical convention. Consider $W = [W_1, W_2, \ldots, W_t, \ldots, W_T]$ the sequence of words including punctuation marks at the borders. After classification W is represented by $C = [C_1, C_2, \ldots, C_t, \ldots, C_T]$, where C_t denotes the category that correspond to W_t. Eight categories were employed. In contrast to English, Spanish morphology is very complex. Particularly, verbs in Spanish have too many conjugations, which in turn makes the accurate classification of every word a difficult task from the automatic implementation point of view.

(P2) The string of categories resulted from step (P1) is divided in windows. The window at instant t, O_t, is defined as $O_t = [C_{t-1}, C_t, C_{t+1}]$. As mentioned above, eight categories were employed, which leads to $8^3 = 512$ possible windows.

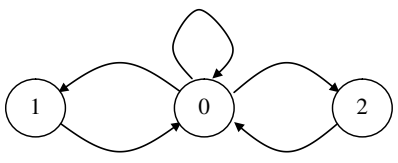

Fig. 2. HMM topology employed to model the intonation discontinuities in Spanish.

Given a observation sequence $O = [O_1, O_2, \ldots, O_t, \ldots, O_T]$ and the HMM parameters $\lambda = (A, B, \pi, \phi)$, the Viterbi algorithm (Huang et al., 1990; Jelinek, 1998) is employed to estimate the optimum sequence of states $S^* = [S_1, S_2, \ldots, S_t, \ldots, S_T]$, where S_t can be 0, 1 or 2 as described above and in Fig. 2.

3.1. Text tagging

The classification procedure described above is based on a POS tagging of the texts. The grammatical tags are required in the training and testing procedures which implies that the tagging must be carried out automatically and efficiently. There is a trade-of between the number of categories used to tag the texts, and the number of parameters that need to be estimated in the observation probability. As a consequence, a very detailed categorization may require a large amount of training data to reliably estimate the output probability. As mentioned above, Spanish has a complex morphology which makes the detailed text tagging a difficult task. To avoid these problems a relatively coarse categorization was chosen in this paper.

A frequency analysis of the words of the training corpus, in accordance to Zipf law, showed that a small subset of the words accounts for a substantial part of the occurrences. It was found that the most frequent occurrences generally belong to a closed set of *grammatical* words (by opposition to *content* words) that can be collected from a static lexicon. Under the assumptions that prosody reflects syntax and that the syntactic outline of a typical sentence can be deduced mainly from the grammatical words, the tagging was implemented by a look-up table that relates words to syntactic or grammatical functions. Words that do not belong to any category in the table are classified as a *non-defined* category. Each word and punctuation mark in the text

string is automatically classified according to one of the following categories:

Class 0: Non-defined category.
Class 1: Punctuation marks that do not define the intonation curve (e.g. comma, colon, semicolon).
Class 2: Punctuation marks that define the intonation curve (e.g. "?", "!", full stop, etc.).
Class 3: Word "que".
Class 4: Demonstrative, possessive and indefinite articles and adjectives (e.g. "el", "los", "un", "una", "mi", "tu", "nuestro", "algún", "cada", etc.).
Class 5: Conjunctions (e.g. "y", "pero", "o", etc.).
Class 6: Non-reflexive pronouns (e.g. "yo", "tú", "él", etc.).
Class 7: Prepositions (e.g. "de", "a", "con", etc.)

It is worth emphasizing that the polyvalent word "que" is considered apart due to its high frequency in Spanish texts. Finally, grammatical word based tagging is less sensitive to misclassification errors and could also be applicable to other languages. For instance, the morphological variability is very high in Spanish. In contrast, although this variability is much lower in English, content word classification may be a significant source of mistakes due to the fact that it depends on the semantic context, which in turn is difficult to model.

3.2. Observation probability estimation with monograms and bigrams

Modelling the observation probability $b_i(O_t) = \text{Prob}(O_t|\text{state } i)$ requires a massive training database as the number of grammatical categories increases. For instance, the parameterization procedure employed here is based on three-word windows. If eight categories are used, the HMM observation probability output is represented by a discrete probability distribution composed of $8^3 = 524$ elements. To counteract this fact, in this paper the output probability was alternatively modelled with monograms and bigrams that reduce the requirements on size of the training database.

3.2.1. Using monograms approximation for the observation probability

Considering that $O_t = [C_{t-1}, C_t, C_{t+1}]$ and assuming that words within a window are independent, $b_i(O_t) = \text{Prob}(O_t|\text{state } i)$ could be approximated as follows:

$$\begin{aligned} b_i(O_t) &= \text{Prob}(C_{t-1}, C_t, C_{t+1}|\text{state } i) \\ &\cong \text{Prob}(C_{t-1}|\text{state } i) \times \text{Prob}(C_t|\text{state } i) \\ &\quad \times \text{Prob}(C_{t+1}|\text{state } i) \end{aligned} \tag{1}$$

where each term of the multiplication can be calculated independently of the others. The estimation of $b_i(O_t)$ according to (1) requires less training data. However, the independence assumption is probably too strong and inaccurate in some cases.

3.2.2. Using bigrams approximation for the observation probability

When compared to monograms, bigrams soften the independence assumption but require more training data. Employing the Markovian approximation, $b_i(O_t)$ could be approximated as follows:

$$\begin{aligned} b_i(O_t) &= \text{Prob}(C_{t-1}, C_t, C_{t+1}|\text{state } i) \\ &\cong \text{Prob}(C_{t-1}|\text{state } i) \\ &\quad \times \text{Prob}(C_{t-1}, C_t|\text{state } i) \\ &\quad \times \text{Prob}(C_t, C_{t+1}|\text{state } i) \end{aligned} \tag{2}$$

3.3. State duration constraints

The probability transitions in HMM are defined by constants that lead to geometric state duration distributions in the ordinary Markov chains. However, the location of the pauses or intonation discontinuities could be related to the length of the *breath* or *intonation groups*. This fact suggests that state duration constraints could improve the accuracy of the prosodic discontinuity estimation. This attempts to model the fact that the speaker's breathing capacity may have an effect on the position of prosodic breaks.

As proposed in (Yoma et al., 2001), in this paper state duration modelling is also included in the Viterbi algorithm by means of the generalization of transition probabilities:

$$a_{i,j}^{\tau} = \text{Prob}(s_{t+1} = j | s_t = s_{t-1} = \cdots = s_{t-\tau+1} = i) \tag{3}$$

where τ is the number of frames in state i up to time t. Consider that $d_{i,j}(\tau)$ is the probability of duration in state i is equal to τ and the following state is j. If $i = 0$, then $j = i = 0$, or $j = 1$ or $j = 2$ given the topology shown in Fig. 2. Notice that state 1 and 2 have duration equal to 1 by definition:

$$d_{1,0}(\tau) = d_{2,0}(\tau) = \delta(\tau) \tag{4}$$

where $\delta(\tau)$ is the Dirac impulse function. As a consequence, the transition probabilities $a^{\tau}_{0,j}$ and $a^{\tau}_{0,0}$ can be estimated by Yoma et al. (2001):

$$a^{\tau}_{0,1} = \frac{d_{0,1}(\tau)}{D_0(\tau)} \tag{5}$$

$$a^{\tau}_{0,2} = \frac{d_{0,2}(\tau)}{D_0(\tau)} \tag{6}$$

$$a^{\tau}_{0,0} = \frac{D_0(\tau) - d_{0,1}(\tau) - d_{0,2}(\tau)}{D_0(\tau)} \tag{7}$$

where $D_0(\tau)$ is the probability of state 0 being active for $t \geqslant \tau$:

$$D_0(\tau) = \sum_{t=\tau}^{\infty} [d_{0,1}(t) + d_{0,2}(t)] \tag{8}$$

From (4) and the topology in Fig. 2, it is easy to show that:

$$a^{\tau}_{1,1} = a^{\tau}_{2,2} = 0, \quad \text{for any } \tau \tag{9}$$

$$a^{\tau}_{1,0} = a^{\tau}_{2,0} = \begin{cases} 1, & \text{if } \tau = 1 \\ 0, & \text{if } \tau \geqslant 2 \end{cases} \tag{10}$$

The state duration distributions $d_{0,1}(\tau)$ and $d_{0,2}(\tau)$ were modelled with the discrete gamma distribution given by

$$d_{0,j}(\tau) = K_{0,j} \cdot \mathrm{e}^{-\alpha_{0,j} \cdot \tau} \cdot \tau^{p_{0,j}-1} \tag{11}$$

where $j = 1$ or $j = 2$, $\alpha_{0,j} > 0$, $p_{0,j} > 0$ and $K_{0,j}$ a normalizing term. The mean duration ($E_{0,j}(\tau)$) and the variance ($\mathrm{Var}_{0,j}(\tau)$) were computed by means of directly observing the training data. The parameters $\alpha_{0,j}$ and $p_{0,j}$ were estimated by

$$\alpha_{0,j} = \frac{E_{0,j}(\tau)}{\mathrm{Var}_{0,j}(\tau)} \tag{12}$$

$$p_{0,j} = \frac{E^2_{0,j}(\tau)}{\mathrm{Var}_{0,j}(\tau)} \tag{13}$$

In contrast to the conditional transition probability in (3), the transition probabilities in the ordinary HMM topology are represented by constants: $a_{0,0}$, $a_{0,1}$ and $a_{0,2}$. Again, observe that $a_{1,0} = 1$, $a_{2,0} = 1$ and $a_{1,1} = a_{2,2} = 0$.

As discussed above, every text segment is broken down into intonation groups. Those segments are always ended by punctuation marks. As a consequence, punctuation marks coincide with pauses and determine the intonation curve in some cases. To model this situation the state duration distributions $d_{0,1}(\tau)$ and $d_{0,2}(\tau)$ were normalized at O_T, where T is the length of the observation sequence, according to:

$$d_{0,1}(\tau) + d_{0,2}(\tau) = 1 \tag{14}$$

3.4. State parameters and punctuation marks

The observation probability distribution in states 1 and 2 is strongly affected by punctuation marks that determine the intonation discontinuity in some cases as mentioned in Section 2.2.1. It was observed that punctuation marks tend to dominate the estimation of $b_i(O_t) = \mathrm{Prob}(O_t \mid \text{state } i.)$ and mask the appearance of intonation discontinuities within the sentence. To overcome this limitation, two output probabilities were defined for every state: one between marks; and, other at the border of a sentence where the end punctuation mark is included in the last observation vector O_T, where T is the length of the sentence.

3.5. Penalizing states 1 and 2 in the Viterbi algorithm

The a priori probability of intonation discontinuities on an inner word is much lower than the a priori probability of not having a pause. To compensate this fact, an additive penalization coefficient, in the log-likelihood domain, was introduced in the Viterbi algorithm to penalize a transition from state 0 to state 0. This coefficient substantially improved the accuracy of the intonation discontinuity HMM estimates.

4. Experiments and results

In order to evaluate the methods presented here, two types of text were employed to test the approach in different cases: 434 and 412 sentences from local newspapers and a theatre play, respectively, were selected. Theatre plays offer a greater variety of prosodic patterns, especially compared to the strong declarative trend of news texts. The segregate analysis based on more than one type of data is not a common practice in the specialized literature. This strategy allows evaluating methods and techniques in different contexts and should lead to more conclusive results.

The following procedure was adopted to annotate the texts: first the texts were recorded by a professional actress who also teaches speech techniques

in a drama school; second, the database was broken down into sentences bounded by punctuation marks; then, the segments were manually analysed, classified and labelled according to the pitch contours shown in Fig. 3. The recorded sentences were processed to extract the fundamental frequency F0 along the utterances by employing "Speech Filing System" (Huckvale, 2001). Then, unvoiced segments were linearly interpolated to obtain a continuous F0 curve. Later, the pitch contour was estimated by band-pass filtering the F0 curve. The filter was designed to reduce the effect of tonal movements at very short and long temporal scales, such as syllable stress and the general falling tendency observed along intonation groups (Prieto, 1998). The main difficulty was to isolate the sought sentence level intonation contours from word stress. This is why a human intervention was needed. The decision of rising or falling intonation was made according to both the pitch curve and a direct perceptual judgement of the putative boundaries. It is worth emphasizing that rather than using a generic, standard and elaborate approach, such as ToBI (Silverman et al., 1992), a straightforward and simplified word aligned markup was applied. The location of each detected final tonal movement is aligned to the word were the phenomenon takes place. As a final step, the intonation employed by the actress was revised and validated by two male speakers that had not pursued any formal training on vocal techniques. Despite this fact, their knowledge about the speech production procedure was enough to allow them to discriminate between rising and falling intonation. The purpose of this revision was to guarantee that the database could be considered representative of ordinary prosodic features. "Vocal technique" denotes here the skill to optimize the use of lungs capacity to increase the breath group length. The validation process is described as follows: first, each sentence was read by both validation speakers in order to perceptually estimate the intonation discontinuities; second, these discontinuities were compared with those obtained by the professional actress. As a result of this final stage, a low percentage of labelling was modified. In fact, it was observed that professional speakers are able to increase the duration of breath groups regardless of punctuation marks.

Each word was associated to a label that indicates the prosodic tonal variation observed at its end (Fig. 3): 0, no tonal discontinuity; 1, falling-tone ending; and 2, raising-tone ending. Cases 1 and 2 indicate that the corresponding word ends an intonation group. It is worth emphasizing that label 0 is never observed at the end of an intonation group.

The newspapers and theatre play data were split into training and testing databases: training, 237 and 202 sentences, respectively; testing, 197 and 210 sentences, respectively. As a consequence two HMMs were trained: one with the newspaper data and the other with the theatre play text. After training, both testing databases were processed by the corresponding HMM based system to evaluate the intonation discontinuities. The estimates were compared to the reference labels and the following rates were computed:

Correct position within segments (CPS): percentage of discontinuities positions correctly estimated within the segments without taking into consideration the type of the discontinuities (i.e. allowing confusions between labels 1 and 2).
Correct type within segments (CTS): percentage of labels 1 and 2 correctly estimated within the segments.
Correct type at the end of segments (CTE): percentage of labels 1 and 2 correctly estimated at the end of the segments.
Correct position (CP): percentage of discontinuities positions correctly estimated considering both inner and end positions, without taking into consideration the type of the discontinuities.
Correct type (CT): percentage of labels 1 and 2 correctly estimated considering both inner and end positions.
Non-discontinuity error within segments (NDES): percentage of labels 0 incorrectly detected as 1 or 2 within the segments.

Text:	.	Voy	a	Ir	a	Santiago	en	la	tarde	.
Categories:	2	0	7	0	7	0	7	4	0	2
Observations:		[2,0,7]	[0,7,0]	[7,0,7]	[0,7,0]	[7,0,7]	[0,7,4]	[7,4,0]	[4,0,2]	
States:		0	0	0	0	2	0	0	1	

Fig. 3. Example of observation and state sequence.

The rates related to inner positions (i.e. CPS, CTS and NDES) depend only on the HMM estimates. On the other hand, CTE strongly depends on the deterministic rules associated to punctuation marks. Finally, CP and CT are a function of both HMM estimates and deterministic rules. It is worth highlighting that the rates that do not take into consideration the misclassification between label 1 and 2 (i.e. CPS, CP and NDES) allow to compare the results presented here with those shown in the literature on the problem of pause detection.

As mentioned above, in this paper HMM is systematically evaluated as a tool to estimate intonation discontinuities and several configurations are tested: C1, corresponds to the baseline system where the observation probability output is directly computed as a discrete probability distribution composed of (No. of grammatical categories)$^{\text{window size}}$ elements as mentioned in Section 3; C2, the observation probability is evaluated with the monogram approximation as described in Section 3.2.1; C3, the observation probability is evaluated with the bigrams approximation as described in Section 3.2.2; C4, indicates that temporal restrictions were introduced in the Viterbi algorithm as explained in Section 3.3; and C5, two output probabilities were defined at state 1 and 2 as described in Section 3.4.

In order to evaluate the optimum penalization coefficient, the objective function

$$E^2 = d \times (1 - \text{CPS})^2 + (1 - d) \times \text{NDES}^2 \qquad (15)$$

was maximized. CPS and NDES are defined above, and d is the percentage of intonation discontinuities in the training texts. In the news text $d = 2.2\%$, and in the theatre play sentences $d = 6.9\%$. This difference in d is certainly a result of the speaker's interpretation skills. It is worth emphasizing that (15) is the weighted Euclidean distance to the optimum point $[(1 - \text{CPS}) = 0; \text{NDES} = 0]$ where the weighting factors are d and $1 - d$. This error measure attempts to give a higher weight to the error $1 - \text{CPS}$ than to NDES. Figs. 4 and 5, obtained with news sentences, show CPS and NDES vs. the penalization coefficient, respectively, with the following combined configurations: C1C4C5; C2C4C5; and, C3C4C5. As can be seen, the lower the penalization coefficient the higher CPS and NDES. Surprisingly, monograms and bigrams estimations of the observation probability, C2 and C3, respectively, give higher CPS (and CTS) than C1 in Fig. 4 when the penalization coefficient is low. This is due to the fact that C2 and C3 are less precise than C1 but they provide better generalization capability. This generalization becomes more important when the penalization increases in module and the influence of the observation probability is

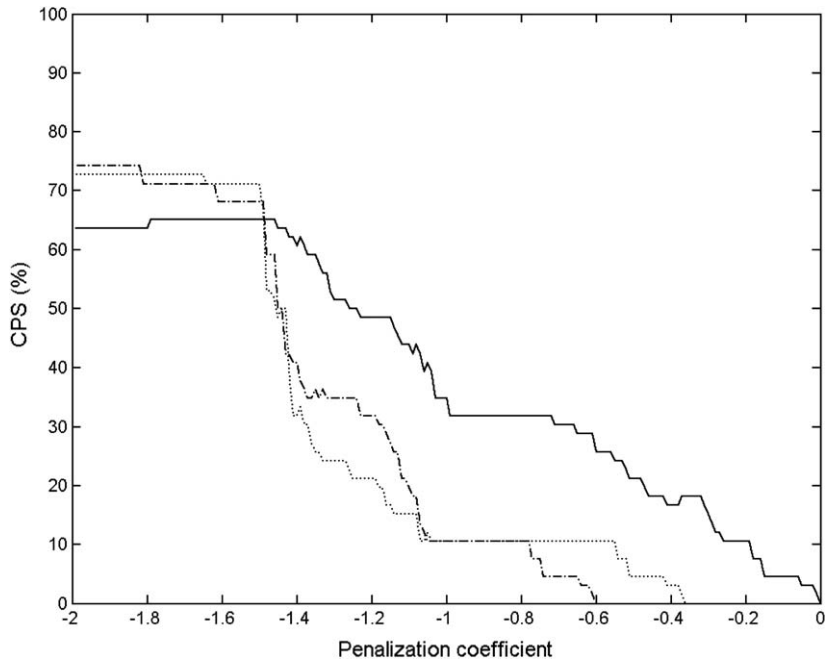

Fig. 4. CPS vs. penalization coefficient in Viterbi with news data: C1C4C5 (solid); C2C4C5 (dotted); and C3C4C5 (dot-dashed).

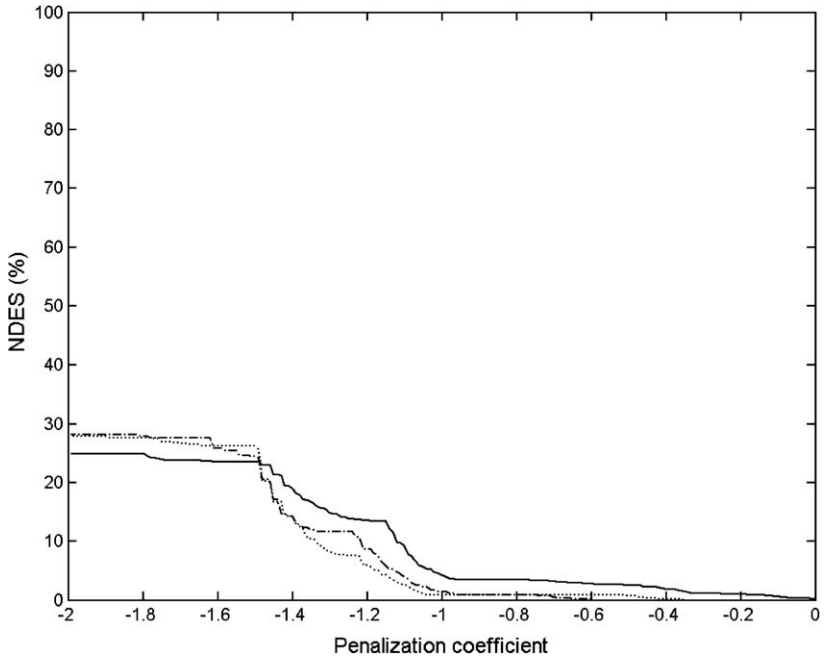

Fig. 5. NDES vs. penalization coefficient in Viterbi with news data: C1C4C5 (solid); C2C4C5 (dotted); and C3C4C5 (dot-dashed).

less relevant. On the other hand, NDES (Fig. 5) also increases because there are more detected discontinuities and the probability of misclassification of a non-pause also increases. However, the increase in CPS is higher than in NDES. This behaviour is emphasized in the sentences from theatre play, as can be seen in the ROC curves of Figs. 6 (news data) and 7 (theatre play data) that present CPS vs. NDES. According to these figures, the difference between C1, C2 and C3 is lower in theatre than in news text.

In the ROC curves shown in Figs. 6 and 7, the best method is the one that provides the dominant curve, i.e. the highest CPS for a given NDES. According to these figures, the observation probability estimation C1 gives usually the lowest discontinuity detection error for a given NDES. Nevertheless, C1 is comparable to C2 and C3 if NDES is higher than 20% in the news sentences and higher than 10% in the theatre play text. Moreover, the improvement due to C1 is less significant in Fig. 7 where the text data corresponds to a theatre play. As mentioned above, C2 and C3 are less precise than C1 but they provide better generalization capability. This generalization becomes more important when the penalization increases in module and more discontinuities tend to be inserted. Due to the fact that the percentage of pauses in theatre play data is higher than in news sentences, this higher generalization capability provided by monogram and bigrams output probability is more noticeable in text from theatre play (i.e. where the speaker introduced more intonation discontinuities).

The C1 vs. C1C4, C2 vs. C2C4 and C3 vs. C3C4 configurations were compared in Tables 2 and 3. According to these results, the introduction of temporal restrictions (C4) was effective to reduce the cost function E defined in (15) with news and theatre play sentences. As can be seen in Tables 2 and 3, the best improvement takes place with C2 (5% or 6% in E) where the observation probability was estimated with monograms. However, when applied in combination with C5, the average reduction in E due to C4 was approximately equal to 7% and 11% with monogram (C2) and bigrams (C3), respectively.

Also in Tables 2 and 3, the effect of C5 on its own is hardly significant. This is due to the fact that employing two observation probabilities to discriminate between intermediate and end observation vectors is not relevant if penalization coefficient is optimized. Notice that every row in Tables 2 and 3 are optimized with respect to this penalization factor. As a consequence, discriminating between intermediate and end vectors enhances the probability of observing an intonation discontinuity. This is the same effect achieved with the penalization coefficient, which in

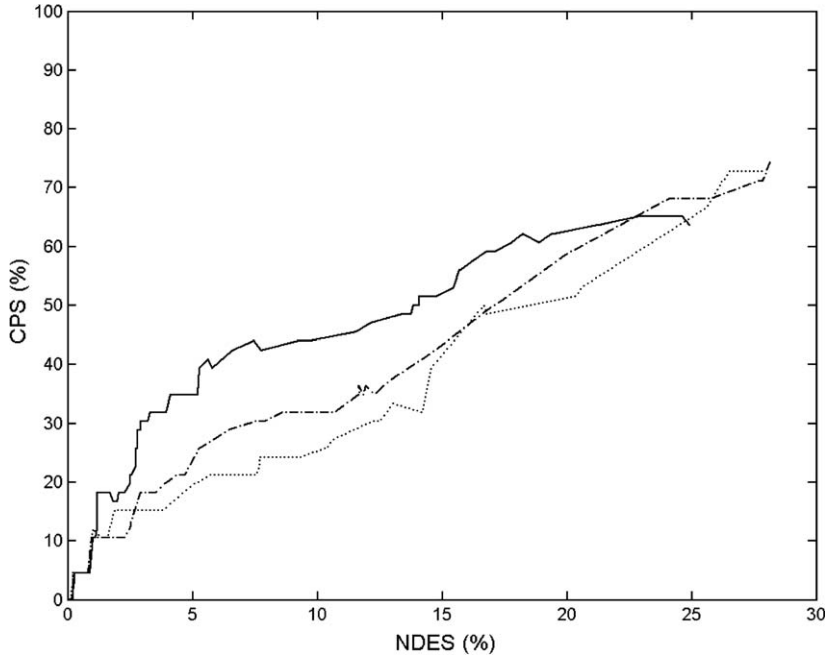

Fig. 6. Receiver operating characteristic (ROC) curves for CPS vs. NDES with news data: C1C4C5 (solid); C2C4C5 (dotted); and C3C4C5 (dot-dashed).

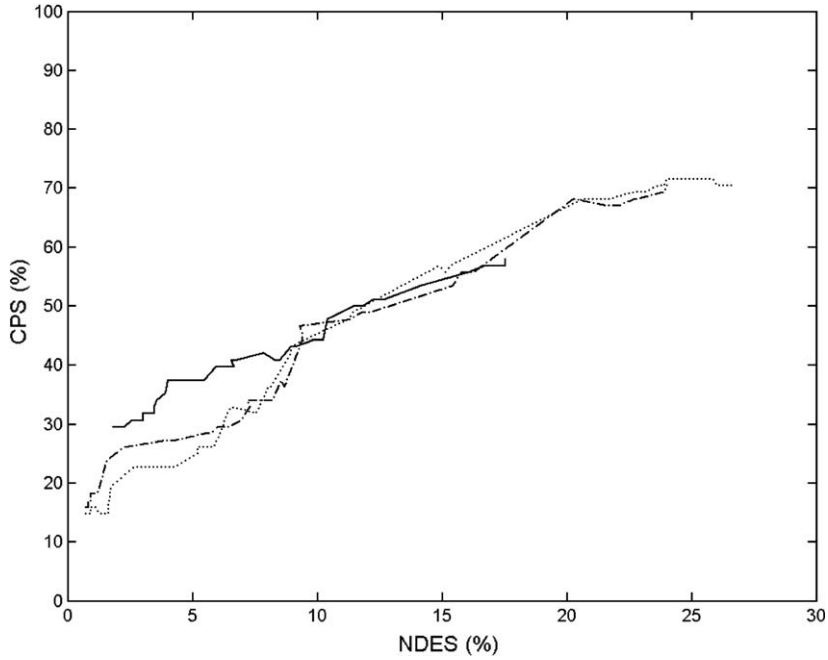

Fig. 7. Receiver operating characteristic (ROC) curves for CPS vs. NDES with theatre play data: C1C4C5 (solid); C2C4C5 (dotted); and C3C4C5 (dot-dashed).

turn ends up masking C5. Finally, with C1, the tendency observed when C4 was applied in combination with C5 is less evident than the improvements achieved with C4 only.

Table 2
Results with news text (%)

Model	CPS (%)	CTS (%)	CTE (%)	CP (%)	CT (%)	NDES (%)	E
c1	31.82	28.79	77.34	82.89	71.10	3.52	0.1078
c2	25.76	21.21	76.56	81.37	68.82	6.96	0.1307
c3	13.64	9.09	76.56	78.33	65.78	3.22	0.1331
c1c4	31.82	30.30	75.00	82.89	70.34	3.37	0.1073
c2c4	21.21	21.21	76.56	80.23	68.82	4.03	0.1245
c3c4	15.15	15.15	71.09	78.71	64.64	3.30	0.1311
c1c5	31.82	30.30	77.34	82.89	71.48	3.52	0.1078
c2c5	21.21	21.21	52.34	80.23	57.03	6.15	0.1327
c3c5	19.70	19.70	53.12	79.85	57.03	4.47	0.1280
c1c4c5	40.91	39.39	75.00	85.17	72.62	5.64	0.1045
c2c4c5	15.15	15.15	71.09	78.71	64.64	1.9	0.1284
c3c4c5	25.76	25.76	71.09	81.37	67.30	5.27	0.1228

Every row is optimized with respect to the penalization coefficient to minimize the error function E.

Table 3
Results with theatre play text (%)

Model	CPS (%)	CTS (%)	CTE (%)	CP (%)	CT (%)	NDES (%)	E
c1	47.73	23.86	77.86	84.41	65.76	10.41	0.1697
c2	28.41	7.95	77.86	78.64	61.02	5.75	0.1954
c3	30.68	7.95	77.86	79.32	61.02	8.13	0.1977
c1c4	44.32	25.00	77.86	83.39	66.10	7.21	0.1615
c2c4	40.91	12.50	63.57	82.37	55.59	10.05	0.1826
c3c4	38.64	12.50	63.57	81.69	55.59	11.96	0.1978
c1c5	43.18	26.14	62.86	83.05	59.32	8.04	0.1677
c2c5	34.09	18.18	62.14	80.34	56.61	9.41	0.1950
c3c5	28.41	15.91	62.86	78.64	56.27	5.02	0.1935
c1c4c5	40.91	20.45	62.86	82.37	57.63	6.58	0.1672
c2c4c5	43.18	19.32	62.86	83.05	57.29	8.95	0.1720
c3c4c5	46.59	19.32	62.86	84.07	57.29	9.32	0.1662

Every row is optimized with respect to the penalization coefficient to minimize the error function E.

The ROC curves that show CTS vs. NDES with C1, C1C4, C1C4, C1C5 and C1C4C5 are shown in Figs. 8 and 9 with news and theatre data, respectively. As can be seen in Fig. 8, C1C4, C1C5 and C1C4C5 are clearly superior to the base line configuration C1. Surprisingly, the combination of C4 and C5 does not lead to better results. This could be due to fact that C4 and C5 are complementary methods and the NDES errors introduced by both approaches are not completely overlapped. This behaviour is even more acute in Fig. 9 where C1C4C5 provided a lower performance than C1C4 and C1C5. Notice that the best result is achieved with C1C4 when NDES is low, although C1C5 is also comparable to C1C4 when NDES increases.

The results presented here suggest that the feature vectors employed by the HMM based system did not provide enough information to reliably predict the type of boundary. The feature vectors depend on the part-of-speech (POS) tagging and

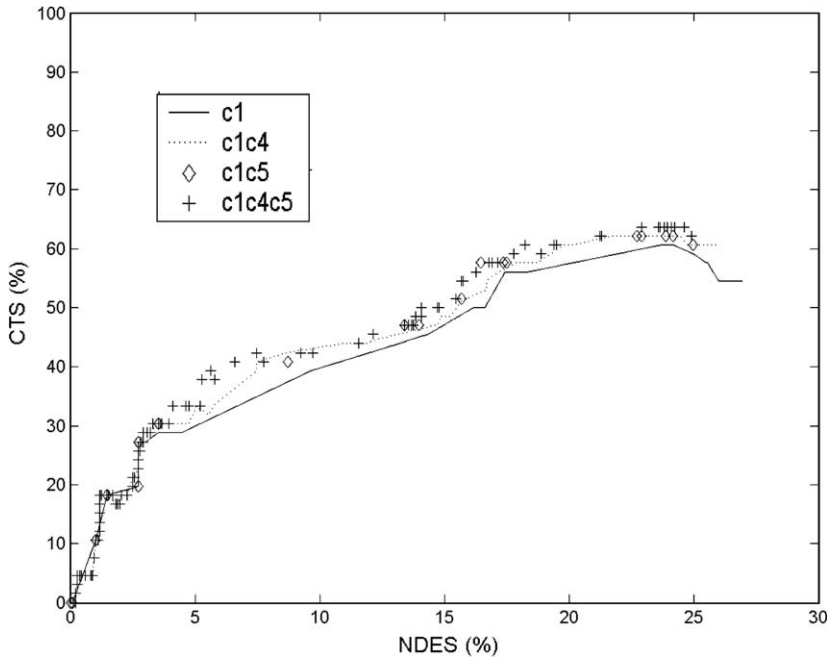

Fig. 8. Receiver operating characteristic (ROC) curves for CTS vs. NDES with news data.

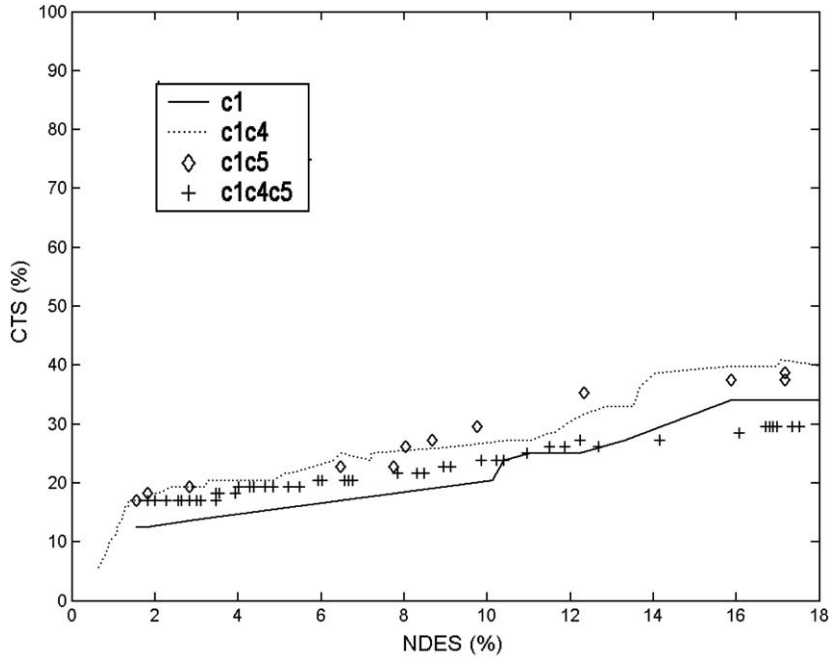

Fig. 9. Receiver operating characteristic (ROC) curves for CTS vs. NDES with theatre play data.

the window size. Adopting a more detailed POS could lead to a higher accuracy in the estimation intonation discontinuity. However, increasing the width of the local analysis window not necessarily should result on a lower estimation error. The size of the training database exponentially increases with the size of the window. This problem is even more severe if a detailed POS tagging is adopted.

5. Conclusions

In this paper the intonation discontinuity detection in Spanish was methodically studied using HMM. A relatively coarse grammatical categorization was employed within two types of texts (sentences from newspapers and a theatre play) read by the same speaker with different styles. The Viterbi-based estimation algorithm also evaluated the type of discontinuity at the boundary of the intonation groups. The HMM observation probability was modelled with monograms and bigrams, besides the ordinary full-window probability, to soften the requirements concerning the database size. State duration modelling was applied to the three-state topology employed here. Also, intermediate and final observation vectors were separately analysed from the observation probability and state duration constraints points of view. Finally, a penalization coefficient in the Viterbi decoding was also evaluated with all the configurations tested in this paper.

The results presented here suggest that state duration constraints can lead to improvements of 5% or 6% in a function error that incorporates the error in discontinuity and non-discontinuity detection. Also, both types of error are strongly affected by the penalization coefficient whose optimal value is dependent on the configurations employed. Moreover, an interaction among the approaches tested here was observed. For instance, the discrimination of intermediate and final observation vectors is usually positive but interferes with the Viterbi penalization coefficient and does not necessarily lead to better results if applied in combination with state duration constraints. The optimal configurations led to reductions of 3% and 5% with news and theatre text, respectively, when compared to the baseline system. Finally, the estimation of the observation probability with monograms and bigrams leads to worse results than the ordinary full-window probability, although they provide better generalization capability. However, the result of monograms and bigrams approximation can be improved when applied in combination with state duration constraints. It is worth emphasizing that monograms and bigrams estimations could be employed to address the

problem of evaluating observation probability of larger windows.

The results show that the type of a prosodic boundary is much more difficult to predict than the position of this discontinuity. This must be due to the fact that the short-term analysis based on the three-element-window used by the HMM system does not provide enough information. A high accuracy estimation of intonation discontinuity would only be achieved by employing a global structure knowledge that can not be determined locally. Particularly, according to Navarro (1944) utterances are divided in tensive and distensive branches with contrasted rising and falling boundary tones. Nevertheless, unravelling such global structure is a task that could be beyond the automatic procedures employed here. Semantic and pragmatic modelling are the missing components that can certainly play a key role in that direction. A relatively better performance was achieved with the newspaper sentences, which can be explained by their more uniform grammatical structure and more neutral reading style.

The task addressed in this paper is very difficult. It is worth emphasizing that the prosodic segmentation provided by human readers is constrained by the syntax of the text. However, actual segmentation also depends on semantic and pragmatic factors, as well as the reader's preferences. Many places where syntactic constraints allow to introduce a non-mandatory segment boundary are not chosen to be emphasized, and the readers' decisions are not necessarily consistent. Consequently, several correct segmentations could be possible depending on the text. Due to the fact that in this paper the prosodic boundaries are estimated only from syntactic information, and validated using a single reading as a reference, the method is prone to generalization errors. The intonation discontinuity detection could be improved by incorporating language modelling that takes into account implicit contextualized semantic knowledge. Finally, the results discussed here could certainly be generalized to other languages.

Acknowledgements

This work was funded by Conicyt/Chile: Fondecyt No. 1030956 and Fondef No. D02I-1089.

References

Agüero, P.D., Bonafonte, A., 2003. Phrase break prediction: a comparative study. XIX Congreso de la Sociedad Española para el procesamiento del Lenguaje Natural. Alcalá de Henares, Spain.

Anderson, M., Pierrehumbert, J., Liberman , M., 1984. Synthesis by rule of English intonation patterns. In: Proc. ICASSP 1, San Diego, pp. 281–284.

Atterer, M., 2002. Assigning prosodic structure for speech synthesis: a rule-based approach. In: Proc. First Internat. Conf. on Speech Prosody SP2002, Aix-en-Provence, France.

Beckman, M., Díaz-Campos, M., Tevis Mcgory, J., Morgan, T.A., 2002. Intonation across Spanish, in the tones and break indices framework. PROBUS, Internat. J. Latin and Romance Linguist. 14 (1), 9–36.

Black, A.W., Taylor, P.A., 1997. Assigning phrase breaks from part-of-speech sequences. In: Eurospeech97, Vol. 2, Rhodes, Greece, pp. 995–998.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks, Pacific Groove.

Fant, L., 1984. Estructura informativa en español. Estudio sintáctico e informativo. Acta Universitis Upsaliensis, Upsala.

Garrido, J.M., Llisterri, J., De la Mota, C., Marín, R., Ríos, A., 1995. Prosodic markers at syntactic boundaries in Spanish. In: Proc. ICPhS Stockholm, Vol. 2, pp. 370–373.

Gussenhoven, C., 2002. Phonology of intonation. Glot Internat. 6 (9/10), 271–284.

Hirschberg, J., Prieto, P., 1996. Training intonational phrasing rules automatically for English and Spanish text-to-speech. Speech Comm. 18 (3), 281–290.

Huang, X.D. et al., 1990. Hidden Markov Models for Speech Recognition. Edinburgh University Press.

Huckvale, M., 2001. Speech Filing System Tools for Speech Research (SFS). University College London. Available from: <http://www.phon.ucl.ac.uk/resource/sfs/>.

Jelinek, F., 1998. Statistical Methods for Speech Recognition. The MIT Press.

Kochanski, G.P., Shih, C., Jing, H., 2003. Prosody modeling with soft templates. Speech Comm. 39 (3–4), 311–352.

Navarro, T., 1944. Manual de entonación española. Guadarrama, Madrid.

Prieto, P., 1998. The scaling of the L tone line in Spanish downstepping contours. J. Phonetics 26, 261–282.

Quilis, A., 1993. Tratado de fonología y fonética españolas. Gredos, Madrid.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., Hirschberg, J., 1992. ToBI: a standard scheme for labeling prosody. In: Internat. Conf. on Spoken Language Processing, Banff, Canada, pp. 867–869.

Sosa, J.M., 1999. La entonación del español. Cátedra, Madrid.

Yoma, N.B., McInnes, F., Jack, M., Stump, S., Ling, L.L., 2001. On including temporal constraints in the Viterbi algorithm for speech recognition in noise. IEEE Trans. Speech Audio Process. 9 (2), 179–182.