Prosodic and Accentual Information for Automatic Speech Recognition

Diego H. Milone, *Student Member*, *IEEE*, and Antonio J. Rubio, *Member*, *IEEE*

Abstract—**Various aspects relating to the human production and perception of speech have gradually been incorporated into automatic speech recognition systems. Nevertheless, the set of speech prosodic features has not yet been used in an explicit way in the recognition process itself. This study presents an analysis of prosody's three most important parameters, namely energy, fundamental frequency and duration, together with a method for incorporating this information into automatic speech recognition. On the basis of a preliminary analysis, a design is proposed for a prosodic feature classifier in which these parameters are associated with orthographic accentuation. Prosodic-accentual features are incorporated in a hidden Markov model recognizer; their theoretical formulation and experimental setup are then presented. Several experiments were conducted to show how the method performs with a Spanish continuous-speech database. Using this approach to process other database subsets, we obtained a word recognition error reduction rate of 28.91%.**

Index Terms—**prosody, accentuation, continuous speech recognition, language models.**

I. Introduction

Considerable progress has been made in automatic speech recognition (ASR) technology over the last 20 years. The incorporation of hidden Markov models (HMM) in ASR in the 1980s led to very high levels of performance, mainly thanks to the way this technique enables the time variability of speech to be modeled. Research using various HMM paradigms have resulted in the incorporation of a number of features that attempt to model human perception. Research has been carried out in the modeling of speech as it relates to the recognition of phonemes, isolated words, connected words and continuous speech (CSR) [52]. In the last few years, techniques such as context dependent phoneme modeling (triphones) and language modeling have been incorporated [23][47].

In recent years, acoustic models have evolved from vector quantization-based models to continuous observation density hidden Markov models (CHMM). In vector quantization systems, acoustic features are modeled as chains composed of a finite set of discrete elements from the vector quantizer. This gave rise to discrete HMM [51]. However, in CHMM it is possible to use continuous observation densities instead of vector quantization, thus taking advantage of modeling speech-selected features through Gaussian mixtures [31]. In the field of acoustic modeling several speech parameterization techniques may be applied, and important advances have been made, such as linear predictive coding [50], cepstral coefficients and mel-cepstral with delta and acceleration coefficients [18].

Computing optimization and its associated algorithms is another field in which considerable progress has been made. Semi-Continuous HMM (SCHMM), also termed tied-mixture models, are examples of algorithms aimed at computational efficiency. The remaining innovations worth mentioning are those related to the speaker's adaptation of estimated parameters and robust speech recognition [25].

Neural networks constitute another important technique that has been successfully applied in some aspects of ASR. In this area, the pioneering works in self-organizing maps [26] and time-delay neural networks [66] should be cited.

Although exhaustive research has been carried out in the field of ASR, computers are still far from attaining the recognition capabilities of human beings [32]. One of the fields in which meaningful improvements have not yet been made is the incorporation of prosodic features into the recognition process. In contrast, we find that prosody is given fundamental importance in text-to-speech (TTS) systems [55]. These analyses and proposed models provide important data about the natural way in which human beings use prosody in spoken discourse. Basically, in the case of TTS systems, prosody gives the naturalness sought in the synthesized speech [61]. Furthermore, some very interesting experiments have studied human speech recognition abilities in different prosodic conditions [21] (also [6] for infants, [27] for spontaneous speech and [35] for dialogue/monologue). A typical situation, encountered daily, is the difficulty in recognizing speech affected by regional accents [2]. This has been studied in the context of ASR in [22]. Another case in which prosodic information and its utilization in spoken language can be seen is speaker identification, see for example [56].

Manuscript received June 2, 2000. This work was supported by the National University of Entre Ríos (UNER-PID#6036&6062), National University of Litoral (UNL-FOMEC#542) and University of Granada.

D. H. Milone is with Cybernetics Laboratory, Department of Mathematics and Informatics, Department of Bioengineering, FI-UNER and Department of Informatics, FICH-UNL. Cybernetics Laboratory: Ruta 11, Km. 10, CP 3101, Oro Verde, Argentine (phone: +54 (343) 4975100 ext 126; fax: +54 (343) 4975978 ext 105; e-mail: d.milone@ieee.org).

A. J. Rubio is with Department of Electronics and Computers Technology, UGR, Spain (e-mail: rubio@ugr.es).

In addition, it is important to note that prosodic modifications in the utterance evidently induce a considerable modification of other parameters that are explicitly modeled in current recognizers. For example, it can clearly be seen that the spectral characteristics of vowels are modified to a significant degree when the intonation changes. Furthermore, the duration of phonemes (mainly vowels) undergoes a notable variation depending on the semantic, syntactic and even orthographic characteristics transmitted in spontaneous speech [4][13]. Considerable improvements in recognition performance can be made by simply taking into account the speaking rate [9]. These, and many other modifications can be achieved at phrase level as well as at word level, or even at syllable and phoneme levels. For example, a fundamental frequency (F_0) model based on several added hierarchic levels is used in [54]. However, prosodic characteristics are often not related to the information at the phonetic level and are only associated with phrase semantics [30].

If we consider prosody within the limits of the word but without arriving at the phoneme, we encounter syllabic suprasegments. These structures are at a superior level than phonemes and are affected by common prosodic features. Some authors (e.g. [49]) treat the terms *suprasegment* and *prosodeme* as synonyms. It should be mentioned that the information provided by the energy and F_0, at suprasegmental level, is not modeled in an explicit way in current recognizers.

These are the principal reasons for which we have sought to address the problem of incorporating prosodic characteristics into ASR in a more explicit way. Some authors [45] have alluded to the potential benefits of incorporating prosodic features into ASR, but they have not proposed concrete solutions. On the other hand, there have been projects that have incorporated some prosodic characteristics to solve a limited range of ASR-related problems. For example, in [33] pitch is used successfully to recover some particular recognition mistakes in connected digits. In this case, as in [17], prosody is incorporated on the basis of post-recognition analysis and not as part of the recognizer itself. In [3], [40], [57], [64] and [67], N-best recognition outputs are used as a starting point and a subsequent rescoring is performed based on prosody. Conversely, in [62] prior segmentation is performed based on prosodic characteristics and the segments are then recognized individually. Exceptions are [28] and [65], where prosody is used to incorporate the phrase-end and other boundary hypotheses into the recognizer itself. In more recent work [42], many aspects of prosody have been incorporated into a speech understanding system. Other authors ([20], [28], [53], [57] among others) have used prosody focused on boundary or disfluency events. In our study, the use of prosody was principally related to suprasegments within the word.

Much research has also been carried out on the information contained in pitch and its utilization in recognizers for several tonal languages (see [15], [16], [20], [28] and [29]). In these languages, there is a very direct relationship between the word meaning and the pronounced tonal cadence [48]. At this point, it is important to emphasize that the use of prosodic features, and the information encoded in them, varies significantly from a tonal language to a non-tonal language and that extrapolations are frequently invalid. Prosody cross analyses between several non-tonal languages can be reviewed, for example, in [11] and [44]. This study was carried out using a Spanish database, and therefore previous analyses of this language ([5][34]) are of great benefit. As an approach to incorporating pitch in ASR, one could add F_0 to the feature vector used in a standard speech recognizer. In our early tests, this addition was performed in a mel-cepstral, energy and delta parameterization, but no improvement was obtained.

An aspect that is closely related to prosody and forms an important part of this study is accentuation. Parameters that characterize prosody in different languages are found to be closely related to accentuation; see for example [13]. This has also been studied for Spanish in [1] and [49]. We developed further the ideas put forward in these studies. For experimental purposes, we propose a system that is divided into two parts: one that associates prosodic characteristics with word accentuation and another that introduces this information into an automatic speech recognizer based on HMM.

The following section discusses the most significant prosodic features, their measurement and relationship to accentuation. This is followed by the description of a set of techniques aimed at establishing an association between the prosody and the accentual properties of Spanish. Section IV contains a proposal for the explicit incorporation of prosody in ASR. The final sections describe a series of experiments carried out, concluding with the analysis of the results achieved.

II. Prosody and Accentuation

When we talk of prosody, we refer mainly to three physical characteristics of spoken language: energy, fundamental frequency (F_0), and duration. These three parameters can be analyzed on a time segment basis ranging from a short frame to complete phrases. However, more relevant information is found at the suprasegmental level; thus, energy, F_0 or the duration of linguistic structures such as syllables are frequently considered.

A. General Aspects of Prosody

Studies have been made from a linguistic viewpoint into how prosody is manifested at different levels of abstraction in spoken language [49]. If we consider, for instance, intonation functions at the linguistic level, it is possible to distinguish functions such as integrative (from words to phrases), distinctive (interrogatory or declaratory statements) and demarcative (as the case of enumeration) functions, as well as functions at the expressive level or even at that of sociolinguistics (for example in regional intonations). All these functions can be considered and incorporated at either the phrase or word levels. They are of particular interest in dialogue systems where the pragmatic meaning of the phrases, in addition to their semantic content, needs to be captured

[58].

From a physical viewpoint, prosody can be defined as the effect that different combinations of energy, F_0 and duration produce on spoken language. Moreover, the pauses related to punctuation and to word endings –as well as sentence and paragraph endings– are often alluded to as part of the field of prosodic features.

As mentioned in the introduction, many aspects of prosody modeling have been discussed extensively in the area of TTS systems. In this area, various models for different languages can be found (for example, [10] for English, [14] for Mandarin Chinese, [43] for German and [63] for French). Basically, these attempt to generate prosodic parameters from written text. However, this task is inverted when prosody is required to assist in ASR. We now need to discover and characterize the prosodic structures in a natural speech utterance in order to incorporate them into the recognition process. In principle, we do not have the text and must attempt to obtain prosodic features from the utterance. Here, the two main facets of the analysis of the problem emerge: (1) obtaining the prosodic features, and (2) incorporating them into ASR. This section examines the relationship between accentuation and prosody, which may be put forward as a conceptual link between these two facets.

The most important case of a suprasegment is the accent (stress accent). Various distinguishing features arise from considering suprasegments as units both different and independent from phonemes. For example, the same phonemes are considered when speaking the /á/ in "papá" (dad) /papA/ and the /a/ of "papa" (potato) /pApa/. However, they are not the same suprasegments relative to the syllable kernel. This difference is expressed in the accentuation.

B. Accentuation

Accentual typology in Spanish (as in English, German, and Italian) is free. That is to say, the accent can be found in any part of the word. This is not the case in Finnish where the accent is always found on the first syllable or in French where it is always found on the last [1]. Furthermore, it is apparent that the accent, on the isolated words, is intimately related to suprasegmental prosodic features (in the case of Spanish). We notice that an accented syllable possesses greater energy, F_0, and duration in its kernel vowel. Therefore, it might be beneficial to carry out an in-depth study of these relationships in continuous speech and their possible application to ASR. However, previous studies in continuous speech indicate the absence of the matches between accented syllable and maximums of energy, F_0, and duration in the kernel vowel. For example, this can be seen in the case of the use of pitch in English in [68]. In the case of Dutch, an automatic classification of accented and non-accented syllables has been achieved, with 72.6% accuracy in the best case [60]. This classification was based on the relationship between prosody and accentuation for continuous speech in telephony.

Accentuation studies in continuous speech carried out in [49] indicate that 36.56% of Spanish words can be considered atonic (no syllables accented), and 90.23% of words in this category being monosyllabic. These words possess distinctive accentual characteristics according to their grammatical function (for example the accentual distinction between the article "el" (the) and the personal pronoun "él" (he)). However, monosyllabic words are not the main interest of this analysis since syllable accentuation cannot be relatively compared within the same word. In these cases, it is necessary to refer it to the sentence. Of the remaining atonic words, several groups of interest have been found, which were analyzed by the author and are incorporated here into our working hypothesis. In the study cited above, the author also related the grammatical function of words to their tonicity. For example, he distinguished between the preposition "para" (for), an atonic word, and the verb in the second person singular "para" (stop), a tonic word. This study, carried out using 20361 words, constitutes an excellent starting point for subsequent designs. Therefore, we have deepened the analysis to reveal some other interesting points that link accentuation and prosody in continuous speech.

C. Relating Prosody and Accentuation in Continuous Speech

An automatic segmentation and the corresponding computation of energy and F_0 curves were performed using a phrase subset of the Spanish database "Albayzin" [12]. This subset, called "minigeo", consisted of 600 phrases pronounced by 6 female and 6 male speakers (i.e., 50 phrases by each speaker) with a vocabulary size of 200 words.

For syllable segmentation, a CHMM alignment based on transcriptions was used. The CHMM was trained with 12 mel-cepstral coefficients with energy and delta coefficients. Each phoneme and the silence was modeled with a 3 state HMM without Gaussian mixtures. Energy curve computation was based on analysis of windowed speech using a frame shift (FS) of 10 ms and a frame width (FW) of 52ms. The same windowing parameters were used to make an estimation of the pitch curve using a cepstral peak detector. Pitch-doubling errors and other spurious peaks were reduced with a median filter and similar rules to those described on [41]. Pitch curves were verified using thousands of sentences. To obtain duration data, the vowel kernel of each syllable was considered, and those made up of diphthongs were taken in account separately. Manual transcriptions of phrases were used to discard monosyllabic words, leaving 2929 analyzed words. The accentual structure of the words was obtained from the text in accordance with orthographic rules[1] and aforementioned considerations regarding grammatical word functions.

First of all, some of the main characteristics of the analyzed phrases should be noted. As a notation for the accentual structures, an 'L' (from *Low*) is used in the case of non-accented syllables and an 'H' (from *High*) for accented syllables. Using this notation, Table I describes the distribution

[1] In Spanish there is a simple set of rules that make it possible to obtain the accentuation from a written word. In English, for example, an accentuation dictionary would be necessary (similarly to the pronunciation dictionary).

of accentual structures in the analyzed data. It is worth recalling that, in Spanish, only one orthographic accent per word is permitted. There are special cases, such as adverbs ending in /-mente/ that have two physical accents but just one orthographic accent (only two words present the /LHLHL/ structure in the analyzed database, and these are not included in Table I) [49]. Table I also includes words considered to be atonic, where orthographic accentuation is not taken as a reference. Table II shows occurrences of different accentuations according to the relative position of accented syllables. Finally, Table III lists occurrences of words according to the number of syllables they contain.

Taking into account the fact that the accented syllable in *isolated* words is characterized as having greater energy, F_0, and duration, an analysis of how these simple rules are satisfied for continuous speech was carried out. These results are given in Table IV. The percentages were calculated using the times that the parameter maxima matched with the orthographic accent, divided by the total number of matches founded. We see from this table that in 17.71% of the cases no maximum value of any of the three parameters coincided with the orthographically accented syllable. However, we also see that the maxima of both energy and duration match the orthographic accentuation on more occasions than do the maxima of F_0. To provide a more detailed analysis, the percentage matches of each parameter for each syllable are shown in Table V. This percentage was calculated by relating the number of matches for each syllable to the total number of words accented on that syllable. The same table presents an average value to indicate how representative each parameter maximum is in relation to accentuation. The matches increased notably in the case of oxytones (accent on the last syllable) and four-syllable proparoxytones (accent on the third syllable from the end).

It was considered of interest to examine the minima, and so an analysis was made of the minimum energy peak, minimum F_0, minimum duration, and all possible combinations between the maxima and minima. In general, this analysis confirmed that accented syllables are characterized by the maxima of energy and duration. However, the correlation found between accented syllables and the F_0 minima was greater than in the case of the F_0 maxima. These results are shown in Table VI. As before (Table V), the matches per accented syllable were considered by taking the F_0 minimum as a reference value. The matches in this case were 66.76, 31.40, 25.06 and 23.80% for the accent in syllables 1, 2, 3 and 4 respectively and with an average of 36.76% (versus 28.97% for the F_0 maxima case).

The prosodic parameters vary significantly when a word pause occurs before or after the word considered [59]. In the case of phrases without important pauses in the middle, it is the first and last words which are affected most. To confirm the influence of this effect, all the previous statistics were recalculated, eliminating all the words located at either end of the phrase from the count. Thus 1984 words were analyzed and, in general, maximum match rates were found to increase by around 10%, in keeping with the conclusions above.

Given the low correlation between intonation and accentuation, a different set of pre-processed intonation signals were tested with an F_0 curve fit by 6th-degree polynomials. The shape of the resulting fit marked the phrase intonation trend, which has a distinctive function (interrogative, affirmative, exclamative, etc.) This fit was subtracted from the original F_0 curve to obtain the difference-by-fit intonation. This new parameter was analyzed according to the method described. However, no meaningful improvements were found.

To continue with this study of intonation trends, a cadence analysis was performed. In this work, the term cadence refers to upward or downward movements of intonation. To obtain a parameter representative of intonation cadences, a slope was fitted onto the time interval of the syllable in question. Three broad groups were considered [19][33][46]: *falling* pitch, with a negative slope, *level* pitch, with a near-zero slope, and *rising* pitch, with a positive slope. These cadences were analyzed for the intonation curve as well as for the difference-by-fit intonation curve (the latter providing no significantly different results). The association between intonation with a positive slope (rising pitch) and accentuation was found to be about 20% closer than the intonation maximum association. Thus, the results of analyses concerning the F_0 minima were also improved on. With this averaged success rate, rising pitch is as valuable as energy, as seen from Table V.

Another interesting series of tests show how these parameters vary depending on the vowel kernel in the syllable. Duration distributions were analyzed; the average duration was found to be between 54.68 ms (for /e/) and 78.45 ms (for /i/). To obtain a more detailed analysis and to link this data with accentuation, the duration averages and standard deviations per accented and non-accented vowel were studied. Fig. 1 gives the results of this analysis. In all cases, accented vowels have a greater average duration. However, it should be noted that the standard deviations are quite large. Although the statistical significance of this parameter is not very high, it still provides some useful information.

The same study was repeated for F_0 and energy, and the results are shown in Figs. 2 and 3. In the case of F_0, four of the five vowels were found to have a higher average for non-accented syllables. For energy it was found that, in two cases, the non-accented vowel has a higher average, although the standard deviations are very large. As energy varies considerably during a sentence, and even more so between phrases, normalization at word level was considered advisable [46]. Fig. 4 shows the results obtained. Although the deviations continue to be large, the averages for the accented syllables can be seen to be higher than those of non-accented syllables in all of the cases. Finally, results for a similar analysis of intonation cadence are shown in Fig. 5. In this case, only the /e/ does not have a rising characteristic (on average) in the accented syllable. The remaining accented syllables are characterized well by a rising pitch.

To summarize the main results of this section, Tables VII and VIII show the confusion matrices relating positions of the most relevant analyzed parameters and orthographic accents. Note that, to simplify these tables, no matching beyond the 4th syllable is included. This compact presentation allows us to look at many aspects of the study together and compare them directly. Here we can select the most representative parameters: Max. E and Max D in Table VII and Rising Pitch in Table VIII. Nevertheless, these simple counts point to the first important conclusion: in continuous speech, the correspondence between orthographic and physical accentuation is mostly lost, thus complicating the task of extracting simple rules to establish such associations. However, this does not mean that no rules can be extracted. Though more complex, other rules, which do relate these characteristics of continuous speech, can be found.

III. Accentual Structure Estimation

In a CSR system assisted by prosodic information, the estimation of a sequence of accentual structures from speech signals is the first task to be performed. We carried out various tests to estimate these sequences on the basis of the little unambiguous information available about associations between prosodic features and accentual structure. The most relevant experiments and results are summarized in Table IX.

The first tests were carried out using only energy and F_0 because obtaining a duration value requires explicit segmentation, at least into syllables. This process is somewhat more complex in real time and independent of the recognition itself. In previous studies ([7] and [8]) an HMM-based classification of pitch movements was carried out for German. In our first tests, HMMs were used as a recognition system for accentual structures. Different models for accented and non-accented syllables were trained and a bigram language model with possible accentual structures (Table I) was built from valid combinations of these basic models. Based on this structure, with an FS of 10 ms and an FW of 25ms, various tests were performed.

As far as parameterization is concerned, the FS and FW of the analysis window were modified and different polynomial degrees (from 3 to 15) for difference-by-fit intonation were tried. Mel-cepstral with delta and acceleration, linear prediction, and spectral coefficients (not shown in Table IX) were also included. For the models, tests using different numbers of states (between 5 and 15) and different mixtures were performed. Accentuation models for each vowel and diphthong were also tested to model the syllables. Concerning the language model, different relative weights for the bigram language model of accentual structures and acoustic models were tested. Comparisons of the relative influences were made using flat grammars (equal probability for all the transitions). By searching for the best characteristic in each aspect, an accentual structure recognition rate of 56.94% was obtained. A sample sentence and its correct and estimated accentual structures are shown in Table X, to clarify these results.

A second series of results were obtained by training static classifiers. The best results were obtained with neural tree networks (NTN) [37][38]. This method combines the structure of decision trees with nodes that separate patterns through a neural classifier (in these tests one self-organizing map per node was used). In this series of tests, only energy and F_0 were initially used. Energy and F_0 related to word maxima were also tested. Syllable and word segmentation were determined using a standard HMM recognizer. For each word, the input pattern contains the maximum values of the prosodic parameters of each syllable in the word. Because pattern length depends on the maximum number of syllables per word in the database, the elements outside the word are set to zero. A recognition rate of up to 85.65% was achieved in the absence of duration and 89.98% when duration was introduced. However, since this NTN-based method is a static classifier, prior syllable segmentation is required. To solve the segmentation problem without using HMM alignment, various tests with a segmentation method based on evolutionary computation are being evaluated [39]. With successful segmentation, it is also possible to consider the use of rising pitch as valuable information for the estimation.

Finally, our current work focuses on improving the estimation of accentual structures from speech utterances. At present, HMMs partially capture the relationship between prosody and accentuation but obtaining linguistic rules from the resulting models is far from straightforward. Nevertheless, as will be discussed in the following sections, the 56.94% estimation for accentual structures is very helpful for CSR. In the subsequent tests on prosodic and accentual information for speech recognition, the results were obtained using the prosodic-accentual structure sequences (PASS) estimated with the best HMM method (HMM-PASS). The prosodic-accentual structure sequences automatically extracted from database transcriptions (T-PASS) were used to obtain the best-case reference.

IV. Prosodic-Accentual Grammar

In this study we worked on the language model in order to incorporate prosodic and accentual features explicitly into the recognition process. A SCHMM-based recognizer was taken as a reference and a language model was then built taking into account the accentual structure of the words in the phrase to be recognized. Thus, recognition was achieved through some PASS estimation, language model penalization (according to the estimated PASS) and recognition itself with the penalized language model. This simplified three-stage structure was created for the purposes of the experiments.

In this section, we will deal mainly with the method of language model penalization since the third stage consists of a standard recognition procedure. The first step, then, is to consider the formalization of the language model.

A. Statistical Language Models

From a fixed and known vocabulary W of Q words, the ordered sequence of m words:

$$\mathbf{w}^m = w_1, w_2, ..., w_m \qquad w_i \in \mathrm{W} \tag{1}$$

is considered for the recognition task. For each word w_i in this sequence its n-order history is defined as:

$$\mathbf{h}_i^n = w_{i-1}, w_{i-2}, ..., w_{i-n+1} \qquad w_i \in \mathrm{W} \tag{2}$$

The language model is approximated using the designated n-gram as:

$$P^n(\mathbf{w}^m) = \prod_{i=1}^{m} P(w_i \mid \mathbf{h}_i^n) \tag{3}$$

The probability that w_i will be spoken, given its history, can be estimated simply from its frequency of occurrence:

$$\hat{P}(w_i \mid \mathbf{h}_i^n) = \frac{C(w_i, \mathbf{h}_i^n)}{C(\mathbf{h}_i^n)} \tag{4}$$

where $C(\mathbf{x})$ is a function that counts the occurrences of a given sequence $\mathbf{x}$ in the training corpus.

However, in practice some word combinations never take place in a given corpus. Therefore, it is necessary to consider n-gram *smoothing*. This technique allows the probabilities for words without a given n-order history to be estimated. Several techniques are useful for n-gram smoothing [24]. A first approximation, for example, is linear interpolation smoothing [52]. Considering a given history $\mathbf{h}^k$ for which $C(\mathbf{h}^k) > 0$ and $0 < k < n-1$:

$$\hat{P}_I(w_i \mid \mathbf{h}_i^k) = \sum_{j=0}^{k} \lambda_j \hat{P}(w_i \mid \mathbf{h}_i^j) \tag{5}$$

$0 < \lambda_j < 1$ and $\sum \lambda_j = 1$ for $j = 0, 1, ..., k$. Where $\mathbf{h}^1$ corresponds to a unigram model and the probability for history $\mathbf{h}^0$ is defined as:

$$\hat{P}(w_i \mid \mathbf{h}_i^0) = 1/Q \tag{6}$$

One of the most frequently used approaches for grammar estimation and smoothing is the *back-off* n-gram [47]. In the present case:

$$\hat{P}_B(w_i \mid \mathbf{h}_i^k) = \begin{cases} \dfrac{C(w_i, \mathbf{h}_i^k) - \vartheta}{C(\mathbf{h}_i^k)} & \text{if } C(w_i, \mathbf{h}_i^k) > 0 \\ \beta(\mathbf{h}_i^k) \hat{P}_B(w_i \mid \mathbf{h}_i^{k-1}) & \text{if } C(w_i, \mathbf{h}_i^k) = 0 \end{cases} \tag{7}$$

where $\vartheta = 0.5$ and:

$$\beta(\mathbf{h}_i^k) = \frac{1 - \sum_{w_i : C(w_i, \mathbf{h}_i^k) > 0} \hat{P}_B(w_i \mid \mathbf{h}_i^k)}{1 - \sum_{w_i : C(w_i, \mathbf{h}_i^k) > 0} \hat{P}_B(w_i \mid \mathbf{h}_i^{k-1})} \tag{8}$$

B. Prosodic-Accentual Structures and Penalization

Let $\mathrm{A} = \{a_1, a_2, ..., a_P\}$ be the set of word accentual classes. The mapping function $g : \mathrm{W} \to \mathrm{A}$ assigns a class $a_i \in \mathrm{A}$ to each word $w_j \in \mathrm{W}$. Within the set A, we consider a distance measure between two accentual structures $\xi(a_i, a_j)$ which assigns a value between 0 and 1 to each pair of accentual structures.

During estimation of the PASS, a sequence like the following is obtained:

$$\hat{\mathbf{a}}^r = \hat{a}_1, \hat{a}_2, ..., \hat{a}_r \qquad \hat{a}_i \in \mathrm{A} \tag{9}$$

The accentual structure of a word can be compared with the estimation, and a link penalization that is proportional to the distance between the two accentual structures $\xi(g(w_i), \hat{a}_i)$ is created. For example, the estimated accentual structure could be $\hat{a}_i = \mathrm{LLH}$ but the word in the hypothesis could be $g(w_i) = g(\text{"estable"}) = \mathrm{LHL}$. Then, a penalization can be introduced based on distance $\xi(\mathrm{LHL}, \mathrm{LLH})$.

However, some problems are encountered in the definition of this penalization function. First, the r elements in the estimated PASS do not necessarily coincide with the m words. Second, word recognition as well as the estimated accentual structure may be wrong. Thus, when defining the penalization function it is necessary to consider this and other particular situations:

$$\varphi_i(w_i, \mathbf{h}_i^n, \hat{\mathbf{a}}^r) = \begin{cases} \gamma_e & \text{if } i > r \\ (\gamma_s - 1)\xi(g(w_i), \hat{a}_i) + 1 & \text{if } i = 1 \vee i = m \\ (\gamma_n - 1)\xi(g(w_i), \hat{a}_i) + 1 & \text{if } C(w_i, \mathbf{h}_i^n) = 0 \\ (\gamma_w - 1)\xi(g(w_i), \hat{a}_i) + 1 & \text{if } C(w_i, \mathbf{h}_i^n) > 0 \end{cases} \tag{10}$$

where each condition is verified excluding those following it, according to the order shown. The motivation for this penalization was the linear rule $\varphi = (\gamma - 1)\xi + 1$: if $\xi = 1$ then $\varphi = \gamma$; if $\xi = 0$ then $\varphi = 1$. The γ constants are adjusted in accordance with the weighting required for each penalization.

The first condition considers the case in which the phrase to be evaluated contains more words than were estimated in the PASS. Secondly, words with a pre- or post-pause are considered. This is necessary since, as we have already seen, the relationship between prosody and accentuation is significantly modified in pre- and post-pause conditions, and therefore estimation is less reliable. Thirdly, the case is considered in which the history $\mathbf{h}_i^n$ (for w_i) is not found during estimation of the initial language model. This probability is the result of a smoothing process. Finally the case is considered where the word probability is calculated from the occurrences counted in the training corpus.

The language model presented in Section IV.A cannot be modified while the recognition process is in progress. Given the word and its history, the probability is defined independently of the speech signal to be decoded and recognized. To incorporate PASS, it is necessary to consider the penalization factor directly in the language model equation in the following way:

$$\hat{P}_P(w_i \mid \mathbf{h}_i^n) = \varphi_i(w_i, \mathbf{h}_i{}^n, \hat{\mathbf{a}}^r)\hat{P}_B(w_i \mid \mathbf{h}_i^n) \tag{11}$$

Now, in the language model, the probability for a given word depends not only on its history but also on its accentual structure. Furthermore, this probability is related to the position of the word in the sentence and to the accentual structure estimated for this position.

For example, let us consider two consecutive words in a simple bigram model. Suppose that we are interested in the probability that the next word will be /rIo/ (river) since the previous was /el/ (the). In the corpus used, we can find the following examples:

1. ***El río** Ebro, ¿pasa por la Comunidad Autónoma de Navarra?* (Does the river Ebro flow through the Autonomous Region of Navarre?)
2. *Mar donde desemboca **el río** Pisuerga*. (The sea into which the River Pisuerga flows)

Let the first sentence be the one uttered and the second sentence another hypothesis in the recognition search. Let us further suppose that the accentual structure was correctly estimated from the prosodic features, that is: /L HL HL HL L L LLLH LHLL L LHL/. The transition probability is certainly not the same when /el/ is the first word and /rIo/ the second in the sentence as when /el/ is the fourth and /rIo/ the fifth in the sentence. This is because the estimated accentual structure in different positions within the phrase is different, and therefore the distance measurement ξ is different. Finally, the penalization function φ makes a different link probability for the same word transition at different phrase positions.

C. *Experimental Setup*

A bigram language model is first calculated using the back-off method. This model subsequently constitutes the reference point for the recognition of all the phrases.

As explained above, for each transition between words, a different probability set must exist depending on the word position in the phrase. To obtain this, a prior grammar net "expansion" is proposed. Fig. 6 shows a simplified scheme of a "compressed" bigram model. In this compressed grammar net, the modification (penalization) of any transition probability would be effective for all the positions of the words in the phrase. Fig. 7 shows the same grammar net expanded. Using this more complex network, transition probabilities can be modified throughout the phrase as the words are pronounced (recognized). Thus, starting from a compressed grammar net, a simple program transforms it into an expanded version. This network is taken as the basis for all the succeeding penalizations (we checked carefully that a standard recognizer has the same performance with both the compressed and the expanded grammar net versions). With the expanded grammar, the penalization function can be applied to the whole sentence according to the estimated PASS.

A number of points need to be taken into account during the expansion of the grammar. In Fig. 6, it can be seen that the compressed grammar net achieves smoothing by means of a null node (empty circle in the figure). Keeping this smoothing in the expanded grammar net is important, though the influence of this on the results obtained must not be overlooked. Each layer in the grammar net corresponds to a word (or silence) and therefore will be expanded up to the maximum number of allowed words per phrase. Hence, this maximum is defined beforehand, reducing recognizer flexibility. However, the main goal of the scheme presented here is to quantify the improvements obtained by the explicit incorporation of prosody into the recognizer. Moreover, as stated above, the three phases into which recognition is divided are only accepted as a compromise for the sake of simplicity and flexibility.

The expanded grammar net has the arcs from the null node of each layer towards silence and from silence towards the null node of the following layer. This allows a silence or pause to be incorporated between two words in the middle of the phrase. Besides, short pauses constitute a model with one active state, which is equal to the central state of the silence. Optionally, each word model has a short pause at the end.

Five basic types of transition can be distinguished in an expanded grammar net. First, those between two words (penalization factor γ_w). Second, those between a null node and a word or vice versa (penalization factor γ_n)(smoothed transitions). Then, those between a silence and a word or vice versa (penalization factor γ_s) and fourth, those which are made towards the terminal node (EXIT) or from the initial node (ENTER towards the first silence with probability 1.0). Lastly, when the phrase of estimated accentual structures is finished, all the transitions of the following layers are affected by the penalization constant γ_e.

The distance measure between accentual classes in set A remains to be defined. For the current tests, this distance was simply based on the Kronecker delta $\xi(a_i, a_j) = 1 - \delta_{i,j}$ for $0 < i, j \le P$. The mapping function $g(w)$ is defined by the orthographic rules of the language used if they exist, or by an accentuation dictionary otherwise. The word accentual classes forming set A are shown in Table I.

Once prosodic-accentual penalization of grammar had been performed, the standard recognition process was carried out using this expanded and penalized grammar.

V. Results

Two basic types of experiments were designed. The first type was intended to investigate prosodic-accentual penalization patterns and the influence of different penalization constants. The second series of experiments was designed to compare the recognition results with the benchmark of a standard HMM-based recognizer. The following section provides details of the corpus, reference system and validation method used.

A. Speech Corpus, Reference System and Cross-validation Tests

The "Albayzin" [12] speech corpus is a continuous read speech database for CSR tasks. This Spanish database consists of 3 corpora of spoken sentences: the *phonetic* corpus (6800 phonetically balanced utterances); the *geographic* corpus (6800 utterances about Spanish geography) and the Lombard corpus (2000 utterances exhibiting the Lombard effect). In our test, the geographic corpus was used, uttered by 134 speakers from central Spain, aged 18 to 55. The environment was a recording studio and the average SNR was 48 dB. The utterances presented a natural spoken language syntax and constrained (geographic) semantics. The average duration of the sentences was 3.55 s. and they comprised 3-26 words per phrase. Some examples of these sentences are:

1. *Dime el número de ríos que desembocan en el Mediterráneo y que sean entre mil y doscientos kilómetros de largo*. (Tell me how many rivers flow into the Mediterranean and are between one thousand and two hundred kilometers long).
2. *¿Cuántos mares hay?* (How many seas are there?).
3. *Quiero saber qué Comunidades Autónomas no tienen salida al mar*. (Please tell me how many Autonomous Regions do not have access to the sea).

Two different subsets from the geographic corpus were used in our tests (one for Section V.B and the other for Section V.C). The first subset (SS1) consisted of 1000 sentences and a vocabulary size of 400 words. SS1 was uttered by 6 female and 6 male speakers. The second subset (SS2) consisted of 600 sentences and a vocabulary size of 200 words. SS2 was uttered by another 6 female and 6 male speakers.

Base-line reference system consists of a SCHMM continuous speech recognizer. The system front-end generates 12 mel-cepstral coefficients with energy and delta coefficients, with cepstral mean subtraction and pre-emphasis filtering. Each phoneme and the silence were modeled with a 3 state HMM and short pauses at the end of words with a 1 state HMM. A standard bigram was used as the language model.

Cross-validation techniques were used to estimate the recognition errors [36]. Of the 1000 sentences in SS1, 600 were used to train acoustic models and to estimate bigram model probabilities, while 400 were used to test the recognizer (Section V.B). In the second set of tests (Section V.C) more detailed cross-validation was used, namely *averaged leave-k-out*. To obtain the overall results, 10 sets of 480 sentences for training and 120 for testing were generated from SS2. Using the training partitions, 10 different sets of acoustic models were obtained and 10 different language models estimated. Finally, from the 10 test partitions, the base-line reference, HMM-PASS and T-PASS results were obtained.

B. Penalization Behavior

Different penalization values were taken for each of the 4 constants. For all combinations of values in the four γ constants, the sentence error rate (SER), the word error rate (WER) (taking into account deletion and substitution errors) and the recognition word accuracy-error rate (WAER) (including also the word insertion errors) were calculated. This exhaustive search was performed using a database subset (SS1) that was different from the one used later for the overall results (Section V.C). The test results for each combination of γ constants were analyzed to determine the penalization behavior.

The combination of penalization constants giving the best results was obtained; this combination was $\gamma_w = -2$, $\gamma_s = -4$, $\gamma_n = -4$ and no γ_e penalization (values expressed as logarithmic probabilities: $\log(p)$). In order to analyze in detail the influence of each penalization constant on recognition rates, we averaged all the results with a given value of the penalization constant. Thus, for example, we averaged all the results with no γ_s penalization to equate this constant influence to zero. By doing this for all the penalization values of γ_s, we obtained a curve that indicated its influence on recognition rates. This curve, together with the corresponding one for the other constants, is shown in Fig. 8 for WER (the corresponding curves for SER and WAER are not shown because their characteristics are similar and they do not provide any other relevant information). Because the curves suggest that $\gamma_s, \gamma_n < -4$ reduces WER, tests for penalizations up to -8 were carried out. For $\gamma_s < -4$, no better results were found. However, for $\gamma_n < -4$ performance went on increasing, even for $\gamma_n < -8$. This reflects the fact that even though the phrases, in both training and test partitions, are always different, both partitions come from the same database and therefore present a similar vocabulary and grammatical structure. Thus, when the smoothing is reduced, part of the grammar net is adjusted to the given database, which reduces overall system flexibility. In order to avoid this effect, it was decided not to surpass the γ_n penalization of -4, and so improvements dependent on a reduction in recognizer flexibility were not obtained. Finally, one further test useful for reference purposes, was carried out; this consisted of eliminating all null nodes (n-gram smoothing) and comparing the recognition achieved with the reference values. This test used the same acoustic models. Under these conditions, the reference model had an error of 4.86% in WER (on 5675 words). The reduction in the WER was 15.43% for the HMM-PASS and 43.00% for the T-PASS. However, these results cannot be considered definitive because n-gram smoothing was eliminated.

C. Overall Results

For these tests, 10 training and test partitions of the SS2 set were used as explained above. With each test partition, the standard recognition results were obtained as base-line reference. The average recognition errors over 10 partitions were 38.30% for SER, 7.54% for WER and 8.53% for WAER.

Grammar nets of language models were expanded and used as a starting point for the tests featuring penalization. For reference, we also tested the recognizer with F_0 added to the speech parameterization and noted a worsening of the results. Table XI shows these and other interesting results. The random penalization is achieved setting random values to ξ in (10). Then we separately apply penalizations of conditions 2, 3 and 4 in (10), also ignoring PASS information.

To obtain definitive results, the best combination of penalization constants found with the SS1 database subset was used. Tests penalizing with HMM-PASS and with T-PASS were performed using continuous speech. Finally, average results over 10 partitions were subtracted from the average reference values and the error reduction rates were obtained by dividing by these reference values. These results are shown in Table XII. This table also provides, as a best reference value, the "overall maximum" results obtained from the best of the 10 partitions and overall combinations of penalization constants.

VI. Discussion and Conclusions

Fig. 8 shows how each of the penalizations used has a different effect. Penalization for the end of sentences, which would seem to be appropriate conceptually, does not produce any improvement in the final results. Penalization for words with an incorrect accentual structure in the null nodes enhances recognition in all cases. This error rate reduction is the same that the presented in Table XI for a fixed γ_n.

The penalization of non-smoothed transitions (γ_w) presents a local error minimum, which allows a more simple selection of its optimum. Finally, the penalization constant that is mainly effective at the beginning and end of a phrase also presents a local penalization minimum but is close to -4. Although not shown in Fig. 8, as mentioned above, similar tests for penalizations of up to -8 have confirmed that, in fact, it is not beneficial to use values for γ_s smaller than -4.

One might expect all penalizations to enhance recognition, even for very high values. However, if excessively high penalizations were used, many paths of the Viterbi search would be discarded and, despite good acoustic scores, an entire hypothesis would be eliminated; obviously, not all accentual structures can be good in HMM-PASS estimation. But even when T-PASS is used, the relationship between the accentual structure and the prosodic parameters is not direct (as discussed in Section II). Thus, we arrive at the same conclusion that it is not advantageous to completely eliminate a hypothesis with an incorrect accentual structure.

During recognition with the expanded grammar, if insertion or elimination errors exist, then all transitions in the following words are penalized. These penalizations punish excessively the insertion and elimination errors. To allow insertions or deletions to take place with a controlled degree of penalization, we tested a grammar net that had been expanded with backward and forward links. However, for recognition rates close to 100%, this type of compensation ceases to have a meaningful effect.

The overall results shown in Table XII clearly indicate the benefits to be obtained from incorporating prosodic accentual information into automatic speech recognition. The table shows the results based on a poor PASS estimate (HMM-PASS). With a better PASS (T-PASS), considerably lower errors are obtained, illustrating how the method operates with a good PASS estimation.

Future projects will be aimed at obtaining improvements in the PASS estimation technique, probably by the inclusion of segmental durations and rising pitch.

Acknowledgment

The authors wish to thank the detailed revision of the editor and reviewers that helped in improving the presentation of this paper.

References

[1] Alarcos Llorach E., *Gramática de la Lengua Española*, Madrid: Espasa Calpe, pp. 52-68, 1999.

[2] Arslan L. M. and Hansen J. H. L., "Language accent classification in American English", *Speech Communication*, Vol. 18, pp. 353-367, 1996.

[3] Bartkova K. and Jouvet D., "Selective prosodic post-processing for improving recognition of French telephone numbers", *Proc. of 7th European Conference on Speech Communication and Technology*, Vol. 1, pp. 267-270, 1999.

[4] Batliner A, Kieβling A., Kompe R., Niemann H. and Nöth E., "Tempo and its Change in Spontaneous Speech", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 2, pp. 763-766, 1997.

[5] Bonafonte A., Esquerra I., Febrer A., Vallverdu F., "A bilingual text-to-speech system in Spanish and Catalan", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 5, pp. 2455-2458, 1997.

[6] Bosch L. and Gallés N., "The role of prosody in infants' native-language discrimination abilities: the case of two phonologically close languages", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 1, pp. 231-234, 1997.

[7] Brindöpke C., Fink G. A. and Kummert F., "A comparative study of HMM-based approaches for the automatic recognition of perceptually relevant aspects of spontaneous German speech melody", *Proc. of 7th European Conference on Speech Communication and Technology*, Vol. 2, pp. 699-702, 1999.

[8] Brindöpke C., Fink G. A., Kummert F. and Sagerer G., "A HMM-based recognition system for perceptive relevant pitch movements of spontaneous German speech", *5th International Conference on Spoken Language Processing*, Prosody and Emotion 6, 1998.

[9] Busdhtein D., "Robust Parametric Modeling of Durations in Hidden Markov Models", *IEEE Trans. On Speech and Audio Processing*, Vol. 4, No. 3, 1996.

[10] Cahn J. E., "A Computational Memory and Processing Model for Prosody", *Proc. of 5th International Conference on Spoken Language Processing*, Prosody and Emotion 2, 1998.

[11] Campione E. and Véronis J., "A Statistical Study of Pitch Target Points in Five Languages", *Proc. of 5th International Conference on Spoken Language Processing*, Prosody and Emotion 5, 1998.

[12] Casacuberta F., García R., Llisterri J. Nadeu C., Prado J. M. and Rubio A., "Development of a Spanish Corpora for the Speech Research", *Workshop on International Co-operation and Standardisation of Speech Databases and Speech I/O Assessment Methods*, CEC DGXIII, ESCA and ESPRIT PROJECT 2589 "SAM", Chiavari, 26-28 September 1991.

[13] Caspers J., "Testing The Meaning of Four Dutch Pitch Accent Types", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 2, pp. 863-866, 1997.

[14] Chen S-H, Hwang S-H and Wang Y-R, "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech", *IEEE Trans. On Speech and Audio Processing*, Vol. 6, No. 3, 1998.

[15] Chiang T-H, Lin Y-C and Su K-Y, "On Jointly Learning the Parameters in a Character Synchronous Integrated Speech and Language Model", *IEEE Trans. On Speech and Audio Processing*, Vol. 4, No. 3, 1996.

[16] Chih-Heng L., Chien-Hsing W., Pei-Yih T. and Hsin-Min W., "Frameworks for recognition of Mandarin syllables with tones using sub-syllabic units", *Speech Communication*, Vol. 18, pp. 175-190, 1996.

[17] Chung G. and Seneff S., "Improvements in Speech Understanding Accuracy Through the Integration of Hierarchical Linguistic, Prosodic, and Phonological Constraints in the Jupiter Domain", *Proc. of 5th International Conference on Spoken Language Processing*, Spoken Language Understanding Systems 1, 1998.

[18] Deller. J. R., Proakis J. G., Hansen J. H., *Discrete-Time Processing of Speech Signals*, Prentice Hall, 1987.

[19] Garrido Almiñana J. M., *Modelización de patrones melódicos del español para la síntesis y el reconocimiento del habla*, Edited by Departament de Filologia Espanyola, Facultad de Filosofia i Lletres, Universitat Autònoma de Barcelona, 1991.

[20] Hirose K. and Iwano K., "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition," *Proc. of IEEE 25rd International Conference on Acoustics, Speech and Signal Processing*, Vol. 3, pp. 1763-1766, 2000.

[21] Hoskins S., "The Prosody of Broad and Narrow Focus in English: Two Experiments", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 2, pp. 791-794, 1997.

[22] Humphries J. J., and Woodland P. C., "The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training", *Proc. of IEEE 23rd International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 317-320, 1998.

[23] Iyer R. M. and Ostendorf M., "Modeling Long Distance Dependence in Language: Topic Mixtures Versus Dynamic Cache Models", *IEEE Trans. On Speech and Audio Processing*, Vol. 7, No. 1, 1999.

[24] Jelinek F., *Statistical Methods for Speech Recognition*, MA: MIT Press, 1999.

[25] Junqua J. C. and Haton J. P., *Robustness In Automatic Speech Recognition: Fundamentals and Applications*, Kluwer Academic Publishers, 1996.

[26] Kohonen T., *The Self-Organizing Map*, NY: Springer-Verlag, 1995.

[27] Laan G.P.M., "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style", *Speech Communication*, Vol. 22, pp. 43-65, 1997.

[28] Lee S-W. and Hirose K. "Dynamic beam-search strategy using prosodic-syntactic information," *Workshop on Automatic Speech Recognition and Understanding*, pp. 189-192, 1999.

[29] Lee T. and Ching C., "Cantonese Syllable Recognition Using Neural Networks", *IEEE Trans. On Speech and Audio Processing*, Vol. 7, No. 4, 1999.

[30] Lieske C., Bos J., Emele M., Gambäck B. and Rupp C. J., "Giving Prosody a Meaning", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 3, pp. 1431-1434, 1997.

[31] Liporace L. A., "Maximum likelihood estimation for multivariate stochastic observations of Markov chains", *IEEE Trans. Information Theory*, IT-28 (5), 1982.

[32] Lippmann R. P., "Speech recognition by machines and humans", *Speech Communication*, Vol. 22, pp. 1-15, 1997.

[33] López E., Caminero J., Cortázar I. and Hernández L., "Improvement on Connected Numbers Recognition Using Prosodic Information", *Proc. of 5th International Conference on Spoken Language Processing*, Prosody and Emotion 2, 1998.

[34] López-Gonzalo E., Rodríguez-García J. M., Hernández-Gómez L. and Villar J. M., "Automatic corpus-based training of rules for prosodic generation in text-to-speech", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 5, pp. 2515-2518, 1997.

[35] Lublinskaja V. and Sappok C., "Speaker attribution of successive utterances: The role of discontinuities in voice characteristics and prosody", *Speech Communication*, Vol. 19, pp. 145-159, 1996.

[36] Michie D., Spiegelhalter D.J., Taylor C.C., *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, University College, London, 1994.

[37] Milone D. H., Sáez J. C., Simón G., Rufiner H. L., "Self-Organizing Neural Tree Networks," *Proc. of 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 20, No. 3, 1998.

[38] Milone D. H., Sáez J. C., Simón G., Rufiner H. L., "Árboles de redes neuronales autoorganizativas," *Revista Mexicana de Ingeniería Biomédica*, Vol. 29, No. 4, 1998.

[39] Milone D. H., Merelo J. J., and Rufiner H. L., "Evolutionary algorithm for speech segmentation," *Proc. of 2002 IEEE World Congress on Evolutionary Computation*, Hawaii, May. 2002.

[40] Molloy L. and Isard S., "Suprasegmental Duration Modeling with Elastic Contraints in Automatic Speech Recognition", *Proc. of 5th International Conference on Spoken Language Processing*, Hidden Markov Model Techniques 3, 1998.

[41] Noll A. M., "Cepstrum Pitch Determination," *J. Acoust. Soc. Am.*, Vol. 41, pp. 293-309, 1967.

[42] Nöth E., Batliner A, Kieβling A., Kompe R. and Niemann H., "Verbmobil: The use of prosody in the Linguistic components of a speech understanding system", *IEEE Trans. On Speech and Audio Processing*, Vol. 8, No. 5, 2000.

[43] Olaszy Gábor and Németh Géza, "Prosody generation for German CTS/TTS systems (from theoretical intonation patterns to practical realisation)", *Speech Communication*, Vol. 21, pp. 37-60, 1997.

[44] Pallier C., Cutler A. and Sebastián-Gallés N., "Prosodic Structure and Phonetic Processing: A Cross-Linguistic Study", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 4, pp. 2131-2134, 1997.

[45] Pols L. C. W., Wang X. and Bosch L. F. M., "Modeling of phone duration (using the TIMIT database) and its potential benefit for ASR", *Speech Communication*, Vol. 19, pp. 161-176, 1996.

[46] Portele T. and Heuft B., "Towards a prominence-based synthesis system", *Speech Communication*, Vol. 21, pp. 61-72, 1997.

[47] Potamianos G. and Jelinek F., "A study of n-gram and decision tree letter language modeling methods", *Speech Communication*, Vol. 24, pp. 171-192, 1998.

[48] Potisuk S., Harper M. P. and Gandour J., "Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method", *IEEE Trans. On Speech and Audio Processing*, Vol. 7, No. 1, 1999.

[49] Quilis A., *Tratado de fonología y fonética españolas*, Madrid: Editorial Gredos, 1993.

[50] Rabiner L. R. and Gold B., *Theory and Application of Digital Signal Processing*, Prentice Hall, 1975.

[51] Rabiner L. R. and Juang B. H., "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, Jan. 1986.

[52] Rabiner L. R. and Juang B. H., *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[53] Rajendran S. and Yenanarayana B., "Word boundary hypothesization for continuous speech in Hindi based on F0 patterns", *Speech Communication*, Vol. 18, pp. 21-46, 1996.

[54] Ross N. K. and Ostendorf M., "A Dynamical System Model for Generating Fundamental Frequency for Speech Synthesis", *IEEE Trans. On Speech and Audio Processing*, Vol. 7, No. 3, 1999.

[55] Rossi M., "Is Syntactic Structure Prosodically Retrievable?", *Proc. of 5th European Conference on Speech Communication and Technology*, Keynote Speech, 1997.

[56] Sönmez M. K., Heck L., Weintraub M., Shriberg E., "A Lognormal Tied Mixture Model of Pitch for Prosody Based Speaker Recognition," *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 3, pp. 1391-1394, 1997.

[57] Stolcke A., Shriberg E., Hakkani-Tür D. and Tür G., "Modeling the prosody of hidden events for improved word recognition," *Proc. of 7th European Conference on Speech Communication and Technology*, Vol. 1, pp. 311-314, 1999.

[58] Swerts M. and Ostendorf M., "Prosodic and lexical indications of discourse structure in human-machine interactions", *Speech Communication*, Vol. 22, pp. 25-41, 1997.

[59] Torres M.I. and Iparraguirre P., "Acoustic parameters for place of articulation identification and classification of Spanish unvoiced stops", *Speech Communication*, Vol. 18, pp. 369-379, 1996.

[60] Van Kuijk and Boves L., "Acoustic characteristics of lexical stress in continuous telephone speech", *Speech Communication*, Vol. 27, pp. 95-111, 1999.

[61] Van Santen J. P. H., "Prosodic Modeling in Text-to-Speech Synthesis", *Proc. of 5th European Conference on Speech Communication and Technology*, Keynote Speech, 1997.

[62] Vereecken H., Vorstermans A., Martens J. P. and Van Coile B., "Improving the Phonetic Annotation by Means of Prosodic Phrasing", *Proc. of 5th European Conference on Speech Communication and Technology*, Vol. 1, pp. 179-182, 1997.

[63] Véronis J., Di Cristo P., Courtois F. and Chaumette C., "A stochastic model of intonation for text-to-speech synthesis", *Speech Communication*, Vol. 26, pp. 233-244, 1998.

[64] Wang C. and Seneff S., "A Study of Tones and Tempo in Continuous Mandarin Digit Strings and Their Application in Telephone Quality Speech Recognition", *Proc. of 5th International Conference on Spoken Language Processing*, Prosody and Emotion 2, 1998.

[65] Warnke V., Gallwitz F., Batliner A., Buckow J., Huber R., Nöth E. and Höthker A., "Integrating Multiple Knowledge Sources for Word Hypotheses Graph Interpretation", *Proc. of 7th European Conference on Speech Communication and Technology*, Vol. 1, pp. 235-238, 1999.

[66] Waibel A., Hanazawa T. Hinton G., Shikano K. And Lang K., "Phoneme recognition using time-delay neural networks", *IEEE Trans. ASSP*, Vol. 37, No. 3, 1989.

[67] Wu S-L., Kingsbury B. E. D., Morgan N., Greenberg S., "Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition", *Proc. of IEEE 23rd International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 721-724, 1998.

[68] Yaeger-Dror M., "Register as a variable in prosodic analysis: The case of the English negative", *Speech Communication*, Vol. 19, pp. 39-60, 1996.

Diego H. Milone (S'95) received the degree in Bioengineering (Hons.) from National University of Entre Ríos, Argentine, in 1998. He is currently finishing his thesis to obtain the Ph.D. degree from Granada University, Spain.

He is with Cybernetics Laboratory, Department of Mathematics and Informatics and Department of Bioengineering, National University of Entre Ríos, Argentine, from 1993, 1995 and 1996, respectively. He is Associate Professor with Department of Informatics, National University of Litoral, Argentine, from 2002.

His current interests research is on signal processing, speech recognition, computational intelligence, bioengineering and auditory prostheses.

Antonio J. Rubio studied physics at the University of Seville, with a specialty in electronics in 1972. He received the Ph.D. degree in 1978.

In 1972 he joined the University of Granada as an Assistant Professor. Since then he has been dedicated to the speech recognition and coding research. He spend a one-year working as a consultant in the Speech Research Department, AT&T Bell Labs, Murray Hill, N. Currently, he is Full Professor in the Department of Electronics and Computer Technology of the University of Granada, Spain and he is the coordinator of the Spanish Network on Speech Technologies.

TABLE I
NUMBER OF DIFFERENT ACCENTUAL STRUCTURES IN ANALYZED DATABASE.

LH	247	LLHL	197
HL	1434	LHLL	220
LLH	186	LLLHL	144
LHL	202	LL	41
HLL	46	LLL	29
LLLH	171	LLLL	12

TABLE II
NUMBER OF DIFFERENT POSITIONS OF THE ORTHOGRAPHIC ACCENT.

Beginning with		Ending with	
H	1480	H	604
LH	671	HL	1977
LLH	383	HLL	266
LLLH	317	HLLL	-

TABLE III
SYLLABLES PER WORD IN ANALYZED DATABASE.

2 syllables	1722
3 syllables	463
4 syllables	600
5 syllables	144

TABLE IV
Matching for Energy (E), Fundamental Frequency (F_0) and Duration (D) Maxima and Orthographic Accent.
✓=Match Between Parameter and Accentuation;
✗=No Match Between Parameter and Accentuation.

Maxima E	Maxima F_0	Maxima D	% of possible combinations
✗	✗	✗	17.71
✗	✗	✓	18.03
✗	✓	✗	4.60
✗	✓	✓	8.19
✓	✗	✗	13.14
✓	✗	✓	17.26
✓	✓	✗	6.34
✓	✓	✓	14.68

TABLE V
Matching for Maxima According to Accented Syllable.
(% of total accented syllables in a given position).

Parameter	Accented syllable 1	2	3	4	Average
Energy	56.55	44.94	25.84	71.38	49.68
Fund. Freq.	42.23	25.44	27.41	20.78	28.97
Duration	70.94	30.35	62.14	53.01	54.11

TABLE VI
Matching of Energy (E) and Duration (D) Maxima, Fundamental Frequency (F_0) Minima and Accentuation.
✓=Match Between Parameter and Accentuation;
✗=No Match Between Parameter and Accentuation.

Max E	Min F_0	Max D	% of possible combinations
✗	✗	✗	11.61
✗	✗	✓	11.82
✗	✓	✗	10.71
✗	✓	✓	14.40
✓	✗	✗	12.07
✓	✗	✓	16.56
✓	✓	✗	7.43
✓	✓	✓	15.38

TABLE VII
Confusion Matrix for Orthographic Accent and Energy (E), Duration (D) and Fundamental Frequency (F_0) Maxima. (Any matching beyond the 4th syllable has been removed to simplify the table)

Number of words		Orthographic accent 1	2	3	4
Max. E	1	**837**	265	**184**	39
	2	618	**302**	32	29
	3	25	75	99	14
	4	0	30	68	**237**
Max. F_0	1	625	**292**	**157**	**146**
	2	**840**	171	25	44
	3	15	111	105	4
	4	0	98	96	69
Max. D	1	**1050**	**270**	69	3
	2	422	204	14	21
	3	8	114	**238**	97
	4	0	84	62	**176**

TABLE VIII
Confusion Matrix for Orthographic Accent and Fundamental Requency (F_0) Minima, Rising Pitch and Falling Pitch. (Any matching beyond the 4th syllable has been removed to simplify the table)

Number of words		Orthographic accent 1	2	3	4
Min. F_0	1	**988**	**283**	92	94
	2	478	211	**154**	39
	3	14	111	96	75
	4	0	67	41	**79**
Rising Pitch	1	**849**	160	66	88
	2	621	**304**	74	22
	3	10	141	**177**	20
	4	0	67	66	**150**
Falling Pitch	1	620	**360**	**157**	76
	2	**836**	176	116	**118**
	3	24	78	59	79
	4	0	57	51	16

TABLE XI
Recognition Errors Rates Using F_0 in Parameterization and Penalizations Without Prosodic-Accentual Information (Average over 10 train/test Partitions).

Recognition Error Rates	SER %	WER %
Base-line ref.	38.30	7.54
MFCC+E+F_0+Δ	39.14	8.08
Random ξ	48.15	10.91
Fixed γ_s	37.79	7.38
Fixed γ_n	31.67	6.29
Fixed γ_w	49.10	12.05

TABLE XII

OVERALL RESULTS FOR PROSODIC-ACCENTUAL AIDED SPEECH RECOGNIZER. BASE-LINE REFERENCE SYSTEM CONSIST OF A SCHMM RECOGNIZER WITH BIGRAM LANGUAGE MODEL. (AVERAGE OVER 10 TRAIN/TEST PARTITIONS).

Recognition Error Rates	SER %	WER %	WAER %	%WER Reduction
Base-line ref.	38.30	7.54	8.53	–
HMM-PASS	30.55	5.36	6.67	28.91
T-PASS	25.50	4.76	5.70	36.87
Overall max.	24.37	3.16	3.79	46.98*

* Error reduction of the best T-PASS partition, relative to the corresponding reference partition.

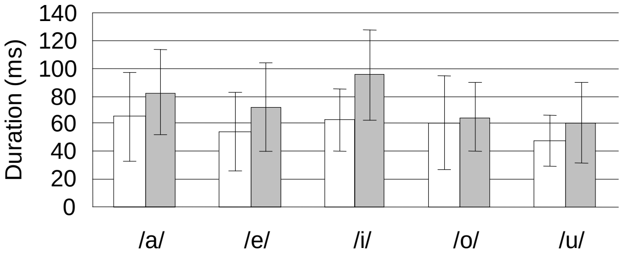

Fig. 1: Duration average per vowel and according to accentuation. Grey bars for accented vowels. Line bars indicate standard deviation.

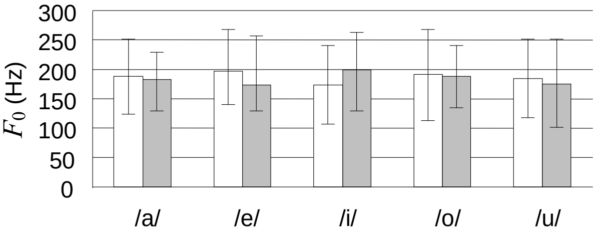

Fig. 2: Fundamental frequency (F_0) average per vowel and according to accentuation. Grey bars for accented vowels. Line bars indicate standard deviation.

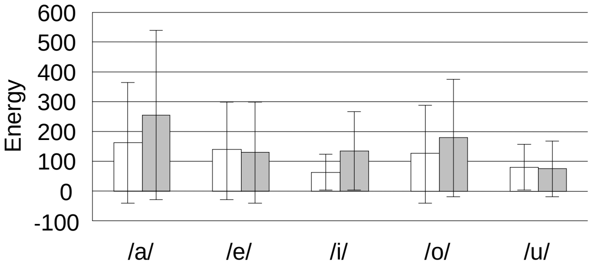

Fig. 3: Energy average per vowel and according to accentuation. Grey bars for accented vowels. Line bars indicate standard deviation. (Values on energy axis are divided by 1E7)

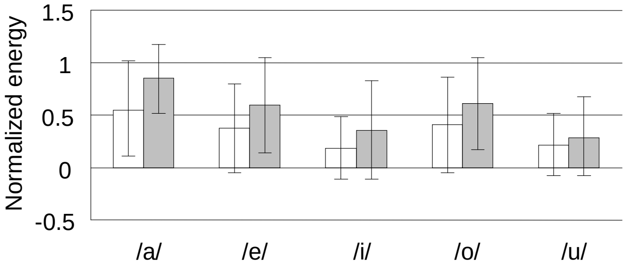

Fig. 4: Word normalised energy average per vowel and according to accentuation. Grey bars for accented vowels. Line bars indicate standard deviation.

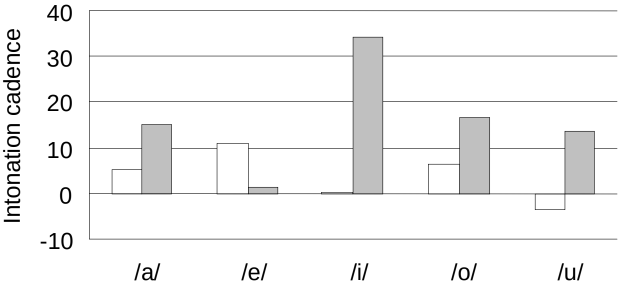

Fig. 5: Intonation cadences average per vowel and according to accentuation. Grey bars for accented vowels.

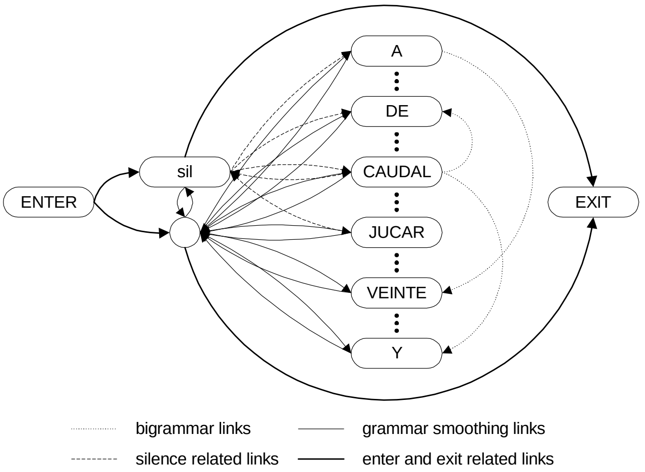

Fig. 6: Standard grammar net. In this case, modifying link probabilities while words are pronounced is not possible.

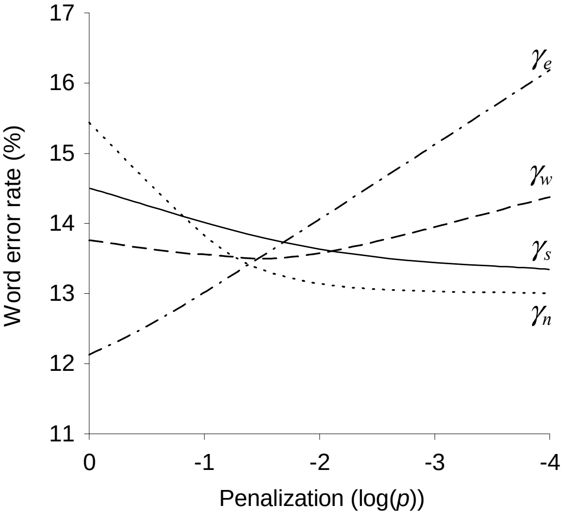

Fig. 8: Influence of different penalization constants.

TABLE IX

BEST RATES FOR ACCENTUAL STRUCTURE ESTIMATION. E = ENERGY; F_0 = FUNDAMENTAL FREQUENCY; D = DURATION; Δ = DELTA COEF.; ΔΔ = ACCELERATION COEF.; DIFF (·) = APPLYING DIFFERENCE BY FIT TECHNIQUE; MFCC= MEL-FREQ. CEPSTRAL COEF.; LPC=LINEAR PREDICTION COEF.; W(·) = NORMALIZING WITH MAXIMA ON EACH WORD; S=STATES PER MODEL; MG=GAUSSIANS PER MIXTURE.

Signal Processing	Rate	Classifier and details
E+ F_0+Δ+ΔΔ	45.56%	HMM, 5 S
E+ F_0+Δ+ΔΔ	53.31%	HMM, 7 S
E+ F_0+Δ+ΔΔ	55.16%	HMM, 7 S, 4 MG
E+diff(F_0)	**56.82%**	HMM, 7 S, obtained with polynomial degree 11
E+diff(F_0)+Δ	50.56%	HMM, 7 S, obtained with polynomial degree 11
E+diff(F_0)+Δ+ΔΔ	50.49%	HMM, 7 S, obtained with polynomial degree 11
diff(E)+diff(F_0)	44.59%	HMM, 7 S, obtained with polynomial degree 13
E+Δ+ΔΔ+MFCC	53.08%	HMM, 7 S
E+Δ+ΔΔ+MFCC	54.88%	HMM, 4 S, 4 MG
E+Δ+ΔΔ+MFCC	43.39%	HMM, 15 S, 4 MG
E+Δ+ΔΔ+MFCC	53.09%	HMM, 5 S, 4 MG, FS 50ms, FW 100ms
E+Δ+ΔΔ+MFCC	**56.94%**	HMM, 7 S, 4 MG, FS 25ms, FW 100ms
E+Δ+ΔΔ+MFCC	50.69%	HMM, 5 S, 4 MG, L and H models each vowel
E+LPC	50.74%	HMM, 7 S, 4 MG, L and H models each vowel
w(E)+w(F_0)	85.65%	NTN + HMM for syllable segmentation
w(E)+w(F_0)+w(D)	89.98%	NTN + HMM for syllable segmentation

TABLE X

EXAMPLE OF HMM ESTIMATED ACCENTUAL STRUCTURE SEQUENCES

Spanish sentence	*Ríos de la Comunidad Autónoma Gallega*
Correct (from orthographic rules)	/HL L L LLLH LHLL LHL/
Estimated with HMM	/H HL L LLLH LHLL L L/
English translation	Rivers of the Galician Self-Governing Territory

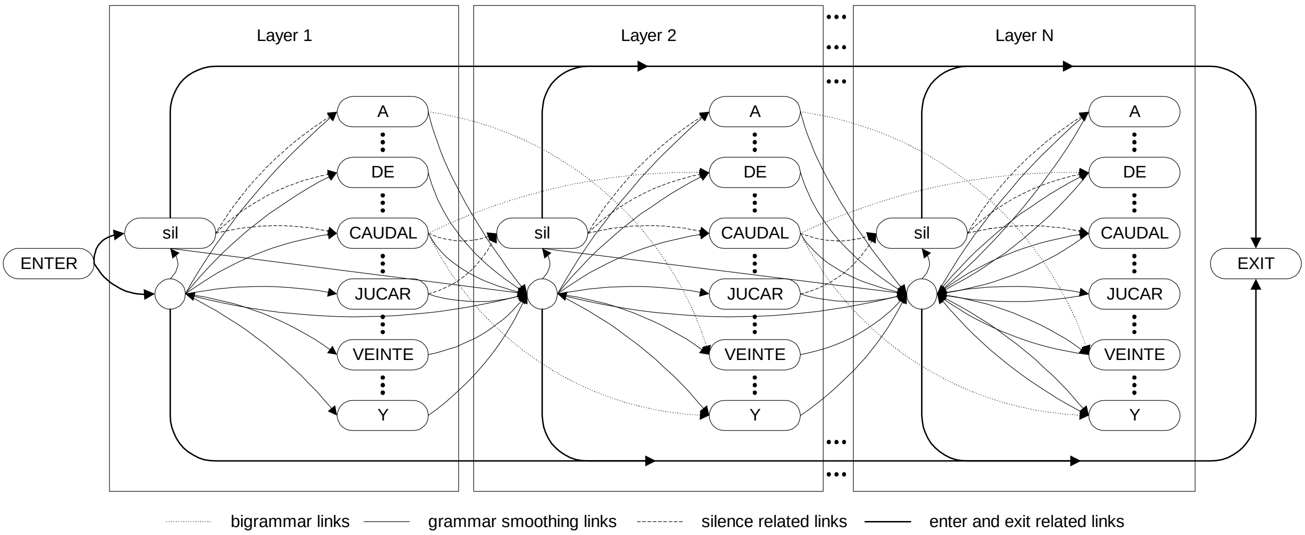

Fig. 7: Expanded grammar net. In this case, alignment between accentual structures of the estimated PASS and layers to perform penalizations is possible.