INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 666-2900 • FAX (510) 666-2956

Automatic Detection of Prosodic Stress in American English Discourse

Rosaria Silipo, Steven Greenberg

TR-00-001

March 2000

Abstract.
Due to the incompletely understood nature of prosodic stress, the implementation of an automatic transcriber is very difficult on the basis of the currently available knowledge. For this reason, a number of data driven approaches are applied to a manually annotated set of files from the OGI English Stories Corpus. The goal of this analysis is twofold. First, it aims to implement an automatic detector of prosodic stress with sufficiently reliable performance. Second, the effectiveness of the acoustic features most commonly proposed in the literature is assessed. That is, the role played by duration, amplitude and fundamental frequency of syllabic nuclei is investigated. Several data-driven algorithms, such as Artificial Neural Networks (ANN), statistical decision trees and fuzzy classification techniques, and a knowledge-based heuristic algorithm are implemented for the automatic transcription of prosodic stress. As reference, two different subsets from the OGI English stories database were hand labeled in terms of prosodic stress by two individuals trained in linguistics. The agreement between the two transcribers on a set of common files is only slightly higher than that obtained by the automatic systems. While the ANN based approach achieves the highest performance (77% primarily stressed vocalic nuclei vs. 79% unstressed vocalic nuclei in average for the two transcribers data sets), the other methods show that both transcribers grant a major role to duration and (to a slightly lesser degree) to amplitude. Pitch relevant features of the syllabic nuclei appear to play a much less important role than amplitude and duration.

1 Introduction

Prosodic stress is an integral component of spoken language [1], particularly for languages such as English that so heavily depend on this parameter for lexical, syntactic, and semantic disambiguation. Therefore, automatic detection of prosodic stress should provide useful information for topic spotting and pronunciation modeling germane to automatic speech recognition.

Experimental and descriptive studies [2, 3] indicate that such prosodic information is mainly based on a complex constellation of information pertaining to the duration, amplitude, and fundamental frequency (pitch) associated with syllabic sequences within an utterance. However, the role played by each one of these acoustic parameters is controversial for American English. For this reason, an automatic detection of prosodic stress is hard to implement, if based only on the currently available knowledge. The traditional perspective attributes the perception of prosodic stress primarely to pitch height and variation [4]. Recent studies focusing on spontaneous speech and on prosodic stress automatic transcription found pitch to be less effective than duration and amplitude [10].

In the current work, several different data-driven approaches are used to implement an automatic transcription of prosodic stress. A set of manually labeled prosodic stress material was created from the OGI English Stories Corpus to train and test the different systems. The goal of this analysis was twofold. First, it sought to implement an automatic detector of prosodic stress with sufficiently reliable performance. Second, the effectiveness of various syllabic acoustic features – duration, amplitude and fundamental frequency – was assessed.

These three basic acoustic features assume very different values across utterances. An investigation of prosodic stress based on the whole syllabic utterance should take into account such differences and provide an adequate normalization to allow meaningful comparisons.

Because a large amount of prosodic stress information is carried by the vocalic nucleus [2, 5] and in order to avoid complicated normalization problems, the role of duration, amplitude and fundamental frequency of only the vocalic nucleus was investigated. Plain, unstressed vowels reasonably produce comparable measures of amplitude, duration and fundamental frequency. In this case an adequate normalization is required only for diphthongs and lengthened vowels.

Some examples of stressed vocalic nuclei in a spoken sentence is shown in figure 1.a), where five vocalic nuclei were marked as stressed by the two transcribers. The last four are characterized by a longer duration and the two diphthongs "ay" also have a higher amplitude. Figure 1.b) shows the corresponding fundamental frequency plot. Here the first stressed vocalic nucleus is characterized by a high constant value of the fundamental frequency.

The automatic transcription system is structured with a pre-processing stage for the isolation of vocalic nuclei and the extraction and normalization of the input features and with a classification stage to distinguish between stressed (S) and unstressed (N) vocalic nuclei.

A number of different data-driven algorithms are implemented for the classification stage of the automatic system. Indeed, the use of different classification strategies makes it possible to determine the best performance, but also provides a means to analyze the problem from different perspectives. The interpretation of the different decision processes might produce interesting insights about the underlying pattern characterizing stressed vocalic nuclei.

In order to obtain an automatic transcription system with sufficiently reliable performance, we first use a Multi-Layer Perceptron to develop a classification algorithm. Artificial Neural Networks (ANN) are by now well known for their capability of constructing extremely accurate decision surfaces on the training set, which usually leads to very good performance on the test set [6]. On the other hand, the interpretation of ANNs' decision process is usually very hard to achieve.

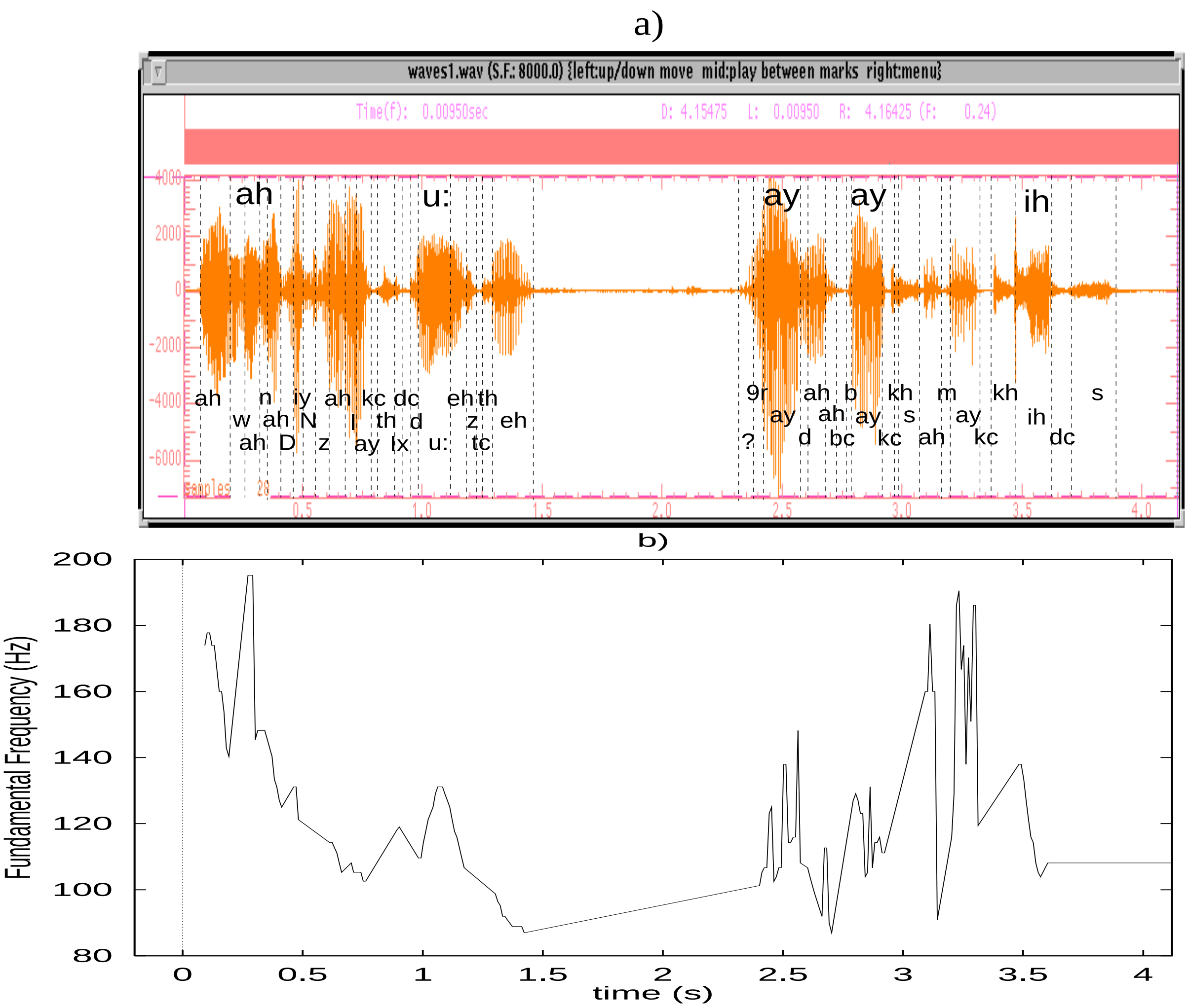

Fig. 1. a) Examples of stressed vocalic nuclei in"uh, one of the things I like to do is to [pause] ride with bikes with my kids". On the bottom is the phonetic segmentation of the spoken sentence and on the top the stressed vocalic nuclei; b) fundamental frequency plot for the spoken sentence in (a)

For this reason, different classification algorithms with a more interpretable structure, such as statistical decision trees [7, 8] and fuzzy systems [9], are also evaluated. The analysis of the structure of the decision trees and of the rules of the fuzzy systems allows us to make certain inferences about the influence of the input features on the final decision process.

Finally, a heuristic algorithm [10] was designed, based on the hypothesis that prosodic stress may be identified as a local maximum of a combination of duration, amplitude and pitch of the vocalic nuclei. Several evidence variables are evaluated to assess the effectiveness of the input features, alone or in combination, in discriminating between stressed (S) and unstressed (N) vocalic nuclei.

As a reference, two different subsets from the OGI English Stories database [11] were manually marked in terms of prosodic stress by two linguistically trained individuals. The agreement between the two transcribers on a set of common files is only slightly higher than what is obtained using the ANN classifier.

2 The Transcribed Corpus

The corpus contains 50-60 second files about a large variety of topics. A phonetic transcription of the files is also supplied. Two different subsets of files are extracted from the database and separately annotated in terms of prosodic stress. The first subset, annotated by transcriber # 1, includes 83 files, with 49 men and 34 women voices. The second subset, annotated by transcriber # 2, contains 51 files, with 38 men and 13 women voices (Table 1). 10 files, 5 men and 5 women voices, are common to both subsets of files, annotated by both transcribers (Table 1). The agreement between the two transcribers on this overlapping part of the two annotated subsets represents one metric of human performance for recognizing prosodic stress and therefore provides a performance benchmark for an automatic transcription system.

Many levels of prosodic stress are usually reported in the literature. However, the concordance among linguists decreases as the number of stress levels increases [12]. It can safely be assumed that only three levels of prosodic stress can be reliably detected by trained linguists: primary stress (S+), absence of stress (N) and an intermediate category with weak stress (S-). Thus, the annotations in the transcribed part of the OGI Corpus refer to primary stressed (S+), to slightly stressed (S-), and to completely unstressed syllables (N).

Table 1. Number of files from the OGI Stories database labeled by each transcriber.

voices	first transcriber	second transcriber	both
men	49	38	5
women	34	13	5
total	83	51	10

2.1 Transcribers Agreement

The agreement between the two transcribers on the common files is shown in Table 2, in order to compare the algorithm's performance with the transcribers' agreement. To simplify the problem of automatic transcription of prosodic stress, only two levels of stress are considered: stressed (S), including primary and minor stress (S+ and S-), and unstressed (N) syllables. The agreement between the two transcribers is evaluated accordingly. Stress labeled as S+ or S- by one transcriber is considered in agreement with the other transcriber, if it was labeled as either S+ or S-.

The first three columns of Table 2 refer to the agreement percentage of transcriber # 1 vs. transcriber # 2, partitioned into utterances spoken by men (M), women (W) and both together (W+M). The second set of three columns refer to the agreement of transcriber # 2 with transcriber # 1. The last three columns refer to the average agreement percentages. The two transcribers roughly agree in recognizing unstressed syllables (N: 84-93%) and primary stress (S+: 90-78%). Much more disagreement occurs in labeling minor stresses (S-: 67-57%). Many syllables marked by transcriber

1 as minor stressed are labeled by transcriber # 2 as primary stressed or unstressed. In general, transcriber # 2 seems to be more biased towards marking primary stress than transcriber # 1. The strongest disagreement about minor stresses regards female speakers.

	Transcr. # 1 vs. # 2			Transcr. # 2 vs. # 1			Average agreement		
	% agreement			% agreement			% agreement		
	S+	S-	N	S+	S-	N	S+	S-	N
W+M	90	67	84	78	57	93	84.0	62.0	88.5
M	93	76	84	81	58	94	87.0	67.0	89.0
W	87	46	85	74	56	92	80.5	51.0	88.5

Table 2. In the first three columns, agreement of transcriber # 1 with transcriber # 2 and, in the second three columns, agreement of transcriber # 2 with transcriber # 1 and in the last three columns the average agreement of the two transcribers on all the common files (W+M), only the male speakers common files (M) and only the female speakers common files (W). S+ primary, S- minor stressed, N unstressed vowels.

3 Extraction of Acoustic Features

Assuming that a phonetic segmentation of the speech file is given, automatic detection of stressed vowels should rely on the analysis of their duration, amplitude and pitch features. In the following subsections, procedures for estimation of fundamental frequency and extraction of acoustic features are described.

3.1 Estimation of Fundamental Frequency

From the original speech signal $x(t)$, n signals $\{\hat{x}_j(t)\}|_{j=1,\dots,n}$ are derived by filtering $x(t)$ over n octaves. The autocorrelation functions, $R_{j,h}(\tau)$, are calculated over a $2N_w$-sample long time window, $\tau \in [h - N_w, h + N_w]$, for each one of the n filtered signals, $\{\hat{x}_j(\tau)\}$. For each time window centered around time h, all of the resulting auto-correlation functions, $R_{j,h}(\tau)$, are summed together across the n frequency channels, as:

$$R_h(\tau) = \sum_{j=1}^{n} R_{j,h}(\tau) \quad (1)$$

If the time window $[h - N_w, h + N_w]$ corresponds to a vowel in the original signal $x(t)$, the total autocorrelation function, $R_h(\tau)$, usually presents a dominant peak at time $\tau_0(h)$, 6-12 ms for male and 3-6 ms for female speakers from the onset of the time window. The fundamental frequency $f_0(h)$ of this time window can be calculated as in eq. 2 [13]. The entire procedure is summarized in figure 2.

$$f_0(h) = \frac{1}{\tau_0(h)} \quad (2)$$

In order to obtain a more robust detection of the peak at time $\tau_0(h)$:

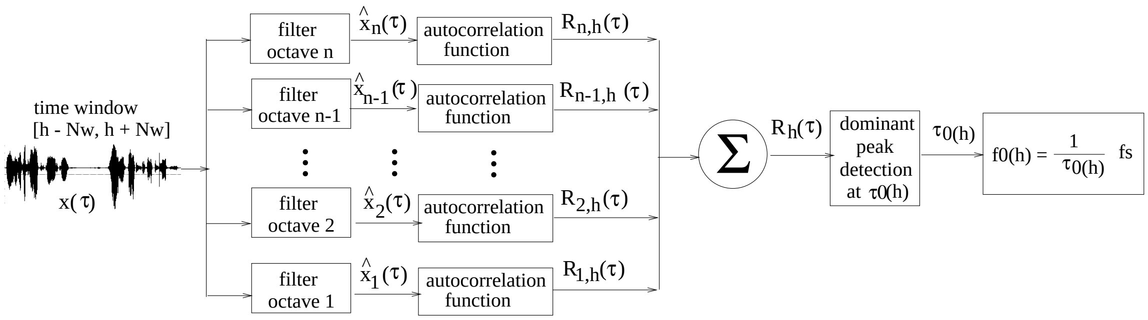

Fig. 2. The procedure for the estimation of fundamental frequency using a time window $2N_w$-sample long and centered around time h.

– a moving average filter is applied to the total autocorrelation function $R_h(\tau)$;
– the signal baseline is approximated by the line joining the average of five $R_h(\tau)$ values at the onset and at the end of the time window, h, and then subtracted from the $R_h(\tau)$ signal.

Generally three situations occur in the detection of the peak at $\tau_0(h)$:

1. only a few peaks are found, indicating a vocalic nucleus (Fig. 3.a);
2. no appreciable peaks are observed, corresponding to a pause or unvoiced consonant (Fig. 3.b);
3. $R_h(\tau)$ presents an oscillatory behavior with peaks of comparable amplitude, corresponding to a voiced consonant (Fig. 3.c).

Since we are interested only in vocalic nuclei, in the latter two cases $f_0(h)$ is not considered. In the first case, the peak with the closest position to the dominant peak in the time window at time $h-1$ is chosen. This technique neutralizes residual outliers reflecting vowel-consonant and consonant-vowel transitions.

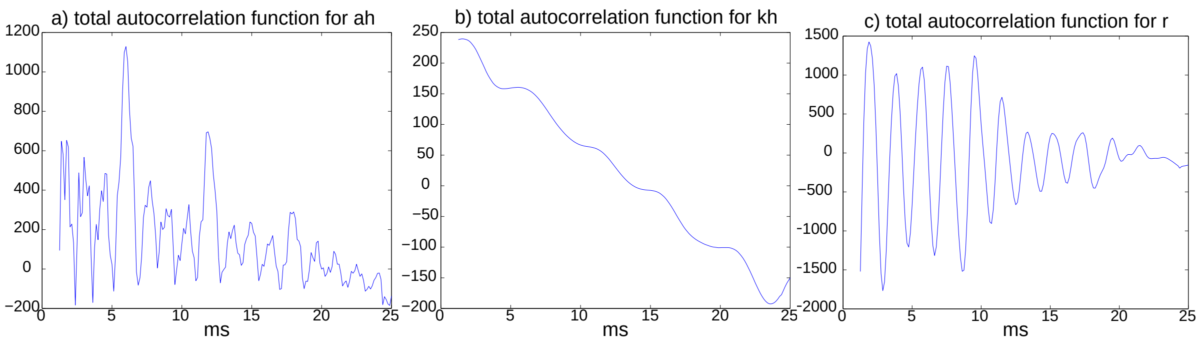

Fig. 3. Autocorrelation functions for short segments of speech, including a) [ah] b) [kh] and c) [r] for a male voice. After 5 ms from the beginning of the window, a dominant peak is clearly identified only for the vowel utterance and it is located at around 6ms. This corresponds to a ca 166.67Hz fundamental frequency.

In this study, 25-ms and 15-ms time windows are considered, both overlapping 5 ms with the previous and following time window. This corresponds to 20 ms and 10 ms time steps, respectively, for the original signal $x(t)$.

3.2 Acoustic Features

For the following analysis, the duration, amplitude and pitch-related features are defined for a generic k-th vocalic nucleus as follows.

1. *Duration* is the number, D_k, of signal samples between its onset and end.
2. *Amplitude*, A_k, is defined as the Root Mean Square of the D_k signal samples $x(t)$ contained in the k-th vocalic nucleus.
$$A_k = \frac{1}{D_k}\sqrt{\sum_{t=1}^{D_k} x^2(t)} \tag{3}$$
3. *Average pitch*, PA_k, refers to the average value of the N_k fundamental frequencies, $f_0(h)$, inside the k-th vocalic nucleus (eq. 4).
$$PA_k = \frac{1}{N_k}\sum_{h=1}^{N_k} f_0(h) \tag{4}$$
4. *Pitch Range*, PR_k, is given as the range covered by the N_k fundamental frequencies $f_0(h)$ in the k-th vocalic nucleus, as:
$$PR_k = \max_{h=1,\ldots,N_k} f_0(h) - \min_{h=1,\ldots,N_k} f_0(h) \tag{5}$$

Every speaker appears to use a different combination of duration, amplitude, average pitch and pitch range, in their pronunciation of stressed vowels. To normalize the variance among speakers, these acoustic features are expressed in terms of variance units from the mean value of their probabilistic distribution estimated inside each utterance.

4 Marking Prosodic Stress

After the pre-processing phase described in the previous section is completed, a classification algorithm is applied. For this task, only two classes – stressed (S) and unstressed syllables (N) – are considered. Finer distinctions among different kinds of stress are not yet taken into account.

The proposed stress assignment procedure focuses on the properties of syllabic vocalic nuclei. Consonants are then discarded before the analysis is performed. Diphthongs, such as [ay] and [oy], that present a longer duration than plain vowels, are divided in two parts. For the same reason, artificially elongated vowels, longer than 250 or 400 ms, are split into three or five parts respectively. The maximum value assumed by each acoustic parameter across all the resulting parts is retained for the analysis. Each acoustic parameter is measured with reference to its average value over the past fifteen vocalic nuclei.

The two subsets of files, annotated by the two transcribers, are evaluated separately. Two thirds of the files of each subset are used to form the training set and the remaining one third forms the test set. In order to give more robustness to the evaluation of the performance of the final system, the training and testing procedure is repeated using a Jack-knifing procedure [14]. The two thirds of the files used as training set and the one third used as test set are cyclically exchanged so that three different pairs of training and test sets are obtained.

4.1 ANN's Classification

In terms of performance of the final transcription system we chose an Artificial Neural Network (ANN) classification paradigm. ANNs have been shown to construct very accurate separation borders among the output classes in training sets and have produced very reliable results in many data analysis fields [6].

In this case, the input vector of the neural network consists of four parameters: duration, amplitude, average pitch and pitch range of each vocalic nucleus. The output layer consists of two neural units, one that is active for stressed (primary or secondary) vocalic nuclei and the other which is active for unstressed vocalic nuclei.

The problem does not present a very high degree of dimensionality either for the input or for the output space and there is not sufficiently reliable knowledge available that could be used to design the ANN architecture. Consequently, the most commonly used ANN paradigm, the MultiLayer Perceptron (MLP), is adopted, consisting of a two-layer, fully connected feedforward architecture and the Back Propagation learning algorithm [6]. In table 3 the performance of an MLP with 2 hidden units is reported as the average percentages of correctly classified vocalic nuclei across the three pairs of training and test sets derived from the Jack-knife method for the two data subsets marked by the two transcribers. The average performance of the system across the two data subsets is shown in the three columns on the right. S+ and S- indicate respectively the percentage of primary and minor stressed vocalic nuclei recognized as stressed by the ANN; N indicates the percentage of unstressed vocalic nuclei correctly recognized as unstressed. Some more experiments are performed, by varying the number of hidden units of the network without obtaining any dramatically different results. The average ANN's performance over the test set is close to the average agreement percentage between the two transcribers reported in the first row of table 2.

	data transcr. # 1			data transcr. # 2			average		
	% correct			% correct			% correct		
	S+	S-	N	S+	S-	N	S+	S-	N
TRAINING	78	54	81	78	55	76	78.0	54.5	78.5
TEST	78	54	81	77	55	77	77.5	54.5	79.0

Table 3. Stressed vs. unstressed discrimination: ANN's performance. S+ primary, S- minor stressed, N unstressed vocalic nuclei.

4.2 Probabilistic Decision Trees

The ANN's average performance over the two data subsets is encouraging and shows the practical feasibility of an automatic detector of prosodic stress. However, the role of the different input features could not be assessed, since the analysis of a neural network's structure is usually a quite complex and time-consuming procedure. Our attention moved to classification algorithms, that are easier to interpret, such as statistical decision trees and in particular its C4.5 implementation for continuous data as described in [8]. Statistical decision trees represent another commonly used data analysis technique. At each step the most informative cut on the input features is performed, where the most informative cut is evaluated in terms of entropy-gain maximization. As a result a more interpretable decision process is generated, because for each step a simple decision (the cut on a given dimension) is made.

The same data configuration used for the ANN is also used for the decision tree. The input vector includes duration, amplitude, average pitch and pitch range of each vocalic nucleus. Two output classes, stressed (S+ and S-) and unstressed (N) vocalic nuclei are considered. The two data subsets are evaluated separately and for each of them a Jack-knifing procedure is applied. The results are shown in table 4 for each transcriber's data set and on average show the decision tree to obtain very similar performance to the ANN.

	data transcr. # 1			data transcr. # 2			average		
	% correct			% correct			% correct		
	S+	S-	N	S+	S-	N	S+	S-	N
TRAINING	78	57	81	82	66	77	80.0	61.5	79.0
TEST	77	54	80	75	57	74	76.0	55.5	77.0

Table 4. Stressed vs. unstressed discrimination: statistical decision tree's performance. S+ primary, S- minor stressed, N unstressed vocalic nuclei.

The structure of the decision tree is shown in figure 4 for one of the three Jack-knife instances of training sets annotated by the first transcriber. The nodes are indicated as circles and the leaves as squares. The letter inside each node indicates the input feature on which the decision is made, the corresponding threshold value is reported on the left and the number of training patterns analyzed by this node is the number on the side of the arrow reaching the node. Inside each leaf the output class attributed to the upcoming training patterns is reported. In this instance no nodes using pitch-related features are found. The decision tree structure for the other two instances are very similar to the one depicted in figure 4, with a few nodes using pitch features added at the bottom layer. The structure of the decision trees constructed on the data labeled by the second transcriber is more complex, producing more than fifty nodes for each training set.

For the sake of clarity a summary of the tree structures for the two sets of data is presented in table 5. The upper rows indicate the percentage of nodes that perform the decision on a given input feature; the rows on the bottom indicate the percentage of training patterns analyzed through a given input feature; that is the percentage of nodes using a given input feature is weighted by the number of training patterns analyzed. From table 5, we can see that the additional complexity in the decision trees constructed on the data labeled by the second transcriber includes a fine-tuning of the classification based on the pitch-related input features. In fact, the number of nodes are roughly equally distributed across the input features, but the number of training patterns that they analyze is not. This means that most of the training patterns are classified by using duration and amplitude and only a few ambiguous training patterns are classified on the basis of pitch-related features.

For both transcribers' data sets, duration and amplitude play a major role in the discrimination of stressed vs. unstressed vocalic nuclei. Pitch-related features are used only to classify a small number of training patterns on the bottom layer of the decision tree.

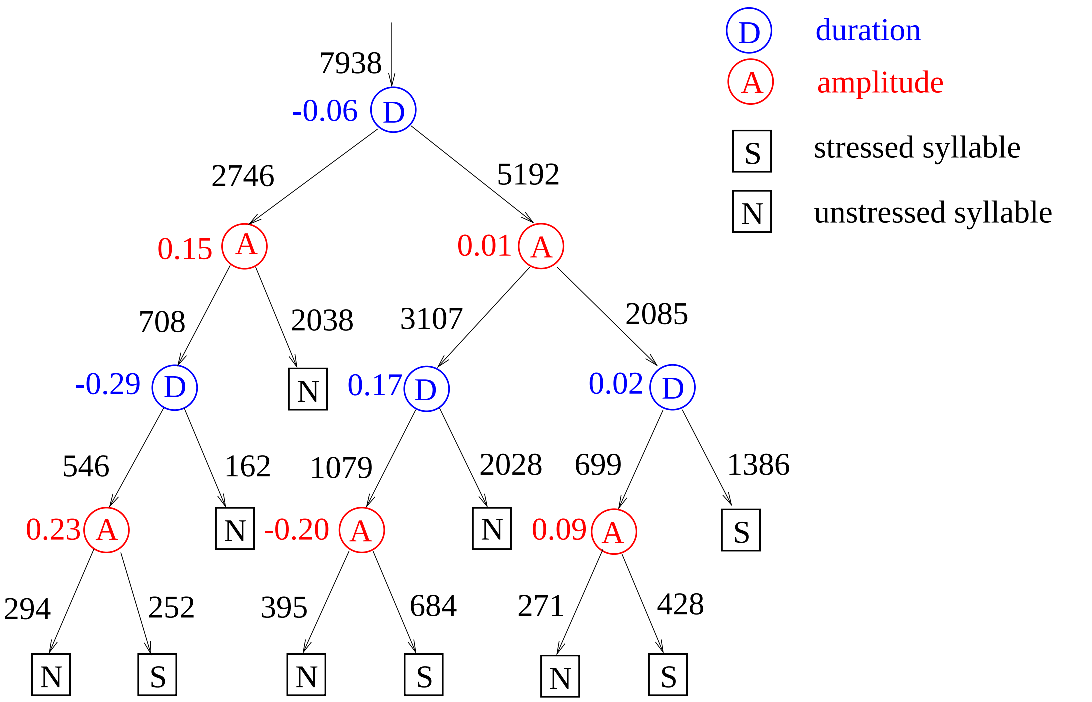

Fig. 4. Decision tree structure for one of the training set instances from the subset of data labeled by the first transcriber. Nodes are indicated as circles and leaves as squares. The letter inside each node indicates the input feature on which the analysis is performed; the corresponding threshold is reported on the left and the number of analyzed training patterns is shown on the side of the arrow reaching it.

	data transcr. # 1				data transcr. # 2			
	% nodes using:				% nodes using:			
	duration	amplitude	average pitch	pitch range	duration	amplitude	average pitch	pitch range
1st training set	42.9	50.0	7.1	0.0	40.3	23.4	18.2	18.2
2nd training set	44.4	55.6	0.0	0.0	34.5	27.6	17.2	20.2
3rd training set	46.2	46.1	0.0	7.7	32.9	31.4	17.1	18.6
average	44.5	50.6	2.4	2.6	35.9	27.5	17.5	19.0
	% training patterns analyzed through nodes using:				% training patterns analyzed through nodes using:			
	duration	amplitude	average pitch	pitch range	duration	amplitude	average pitch	pitch range
1st training set	51.6	47.6	0.8	0.0	52.7	41.3	3.9	2.1
2nd training set	57.4	42.6	0.0	0.0	39.0	49.8	5.1	6.2
3rd training set	50.5	48.8	0.0	0.7	49.3	43.2	2.6	4.9
average	53.2	46.3	0.3	0.2	47.0	44.8	3.9	4.4

Table 5. Summary of the structure of the decision trees constructed on each instances of training set and for each one of the data subsets labeled by the two transcribers and on average.

4.3 Fuzzy Classification

Another classification paradigm usually appreciated for its easy interpretability is based on fuzzy logic. Fuzzy logic derives from a qualitative rather than a statistical characterization of the involved variables and output classes [9]. Nowadays, several automatic algorithms constructing fuzzy rules from a set of examples are available [15] [16]. Even though the performance of fuzzy systems usually do not compare with the ones yielded by ANNs, some insights about the underlying system can come from the interpretation of fuzzy rules. For this reason, an investigation of a possible fuzzy automatic labeling of prosodic stress has also been performed and, for this purpose, the algorithm described in [16] was adopted.

The performance. As in the previous analysis, two thirds of the annotated files for each subset of data are used as a training set to implement a fuzzy model [16], that discriminates stressed (S) from unstressed (N) vocalic nuclei. The resulting fuzzy model is tested on the remaining third of material. Even here, the training and testing procedure is repeated using a Jack-knife method and the average performance of the fuzzy algorithm across the three instances of training and test sets are shown in table 6 for both transcribers' data sets.

	data transcr. # 1			data transcr. # 2			average		
	% correct			% correct			% correct		
	S+	S-	N	S+	S-	N	S+	S-	N
TRAINING	100	100	100	100	100	100	100.0	100.0	100.0
TEST	71	53	77	60	40	80	65.5	46.5	78.5

Table 6. Stressed vs. unstressed discrimination: fuzzy-logic-based algorithm performance. S+ primary, S- minor stressed, N unstressed vocalic nuclei.

The system's performance on the test set drops down a few percent with respect to the ANN's performance, mainly for the recognition of stressed vocalic nuclei. Let us try now to interpret the sets of fuzzy rules generated by the system on the three different instances of the training set for the two data subsets. Due to the problem's complexity, the fuzzy algorithm generates about a thousand rules, which makes a detailed visual analysis prohibitive.

We are mainly interested in the discriminative power that the proposed fuzzy classification yields for each input feature. It is possible to automatically calculate the fuzzy information gain that derives from the use of a given input feature in a fuzzy model [18]. Such an information gain can be used to quantify the discriminative power of each input feature in the fuzzy model.

The information gain. Given a number m of output classes C_i, $i = 1, \ldots, m$, and an n-dimensional input space, a fuzzy algorithm derives a set of N_R fuzzy rules $\{R_k(C_i)\}$ - with $k = 1, \ldots, Q_i$, $i = 1, \ldots, m$ and $\sum_i Qi = N_R$ - mapping the n-dimensional input into the m-dimensional output space. Each input pattern $\mathbf{x} = [x_1, \ldots, x_n]^T$ is associated with each output class C_i by means of a membership value $\mu^k_{C_i}(\mathbf{x})$. Let us consider $Q_i = 1$ for purposes of illustration.

The membership functions $\mu^k_{C_i}(\mathbf{x})$ quantify the degree of membership of input pattern $\mathbf{x}$ to output class C_i. The average degree of membership, $V(C_i)$, of patterns $\mathbf{x}$ to output class C_i across the input space $D \subset \mathcal{R}^n$ is given in eq. 6.

$$V(C_i) = \frac{\int_{D\subset\mathcal{R}^n} \mu_{C_i}(\mathbf{x})\, d\mathbf{x}}{\int_{D\subset\mathcal{R}^n} d\mathbf{x}} \tag{6}$$

The usual information functions, such as entropy or the Gini function [18], could be applied to $V(C_i)$ to quantify the amount of information contained in the system. Such information functions, however, require variables summing up to one across the output classes, which is not necessarily true for the average membership degree $V(C_i)$. To solve this problem, the relative average membership function, $v(C_i)$, of output class C_i can be used, as described in eq. 7.

$$v(C_i) \;=\; \frac{V(C_i)}{\sum_{j=1}^{m} V(C_j)} \tag{7}$$

If trapezoids are adopted as membership functions, the relative average membership function, $v(C_i)$,of each output class C_i becomes particularly simple to calculate [17]. A measure of the information, $I(C)$, contained in the fuzzy model, can be obtained by applying the traditional information functions to the variables $v(C_i)$ (eq. 8).

$$I(C) \;=\; -\sum_{i=1}^{m} v(C_i)\, \log_2(v(C_i)) \tag{8}$$

At this point, a criterion is necessary to quantify how much of the information $I(C)$ contained in the model is exploited by each input feature x_j for classification purposes. In a fuzzy model, each input dimension x_j consists of a number, F_j, of linguistic values, L_k. To classify along input dimension x_j means to define F_j fuzzy sub-models, one for each linguistic value $x_j = L_k$. The information still available in the fuzzy model, after the fragmentation along input feature x_j, is given by the average - $I(C|x_j)$ - of the information still available in all sub-models - $I(C|xj = Lk)$ - as defined in eq. 9 and described in figure 5.

$$I(C|x_j) = \frac{1}{F_j}\sum_{k=1}^{F_j} I(C|x_j = L_k) \tag{9}$$

Thus the information gain $g(C|x_j)$ associated with the use of input feature x_j can be expressed as the relative difference between the intrinsic information available in the system before - $I(C)$ - and after using the variable x_j for the analysis - $I(C|x_j)$ - as in eq. 10.

$$g(C|x_j) = \frac{I(C) - I(C|x_j)}{I(C)} \tag{10}$$

The proposed information gain $g(C|x_j)$ expresses the effectiveness of parameter x_j in performing the required classification on the basis of the given fuzzy rules and therefore can be adopted as a feature-merit measure of input parameter x_j. The less effective the input feature x_j is in the original set of fuzzy rules, the closer the remaining $I(C|x_j)$ is to the original information $I(C)$ and the lower the corresponding information gain is. The input features x_j with the highest information gain are the most used by the fuzzy model to represent the training set and therefore are the most effective for the proposed analysis.

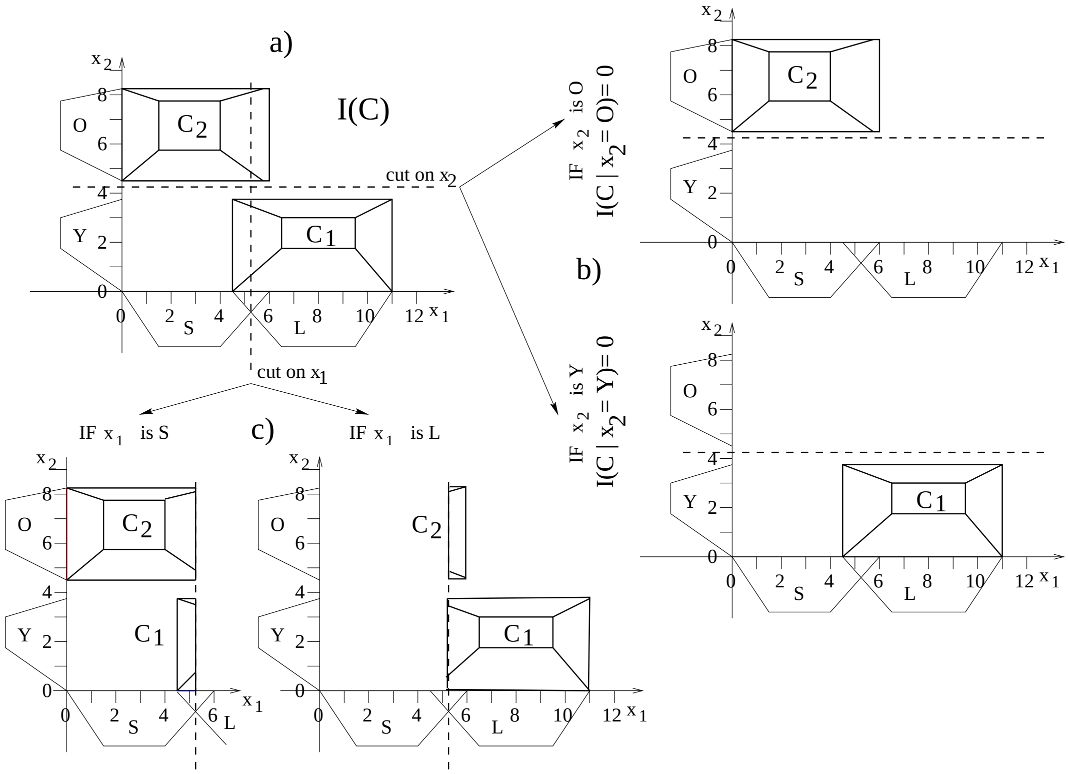

Fig. 5. a) Example of a two-class fuzzy model on a two-dimensional input space. b) and c) sub-models generated by cutting the original fuzzy model in a) along input feature b) x_2 and c) x_1.

The discriminative power of the acoustic features in the fuzzy models. This analysis is applied here to assess the role of the different input features in the implemented fuzzy classification process. In table 7 the information gain related to the use of the input features is reported for the systems with the average performance shown in Table 6. An information gain of 1.0 indicates perfect discriminability of the output classes along this input feature; linearly decreasing values of the information gain (eq. 10) describe increasingly overlapping output classes on the analyzed input feature; an information gain of 0.0 indicates that this input feature is not used by the classifier.

Table 7 shows that the fuzzy system does not use much pitch information when classifying the data labeled by the first transcriber. Indeed the information gain associated with duration and amplitude are dominant in comparison to those pertaining to the pitch-related features. Pitch range gains some importance for one of the three Jack-knife training sets for the data labeled by the second transcriber. It should be mentioned, however, that pitch range and duration are related, because a vocalic nucleus has to be adequately long to allow a high variation of pitch. In general pitch related features, and particularly average pitch, do not seem to be very important for the fuzzy transcriber of prosodic stress for both data subsets, which agrees with the analysis of the decision trees structure described in the previous subsection.

A similar evaluation is conducted to discriminate between primary stressed (S+) and unstressed (N) vocalic nuclei and between minor stressed (S-) and unstressed (N) vocalic nuclei. The corre-

Table 7. Information gain of the input features in the fuzzy models implementing a stressed vs. unstressed vocalic nuclei discrimination for the two transcribers data sets.

	data transcr. # 1				data transcr. # 2			
	information gain for:				information gain for:			
S vs. N	duration	amplitude	average pitch	pitch range	duration	amplitude	average pitch	pitch range
1st training set	0.09	0.21	0.03	0.03	0.11	0.08	0.04	0.08
2nd training set	0.11	0.11	0.02	0.01	0.20	0.12	0.03	0.27
3rd training set	0.10	0.09	0.01	0.01	0.20	0.23	0.04	0.04
average	0.10	0.14	0.02	0.02	0.17	0.14	0.04	0.13

sponding average performance and information gain are reported in table 8 and 9 respectively.

As it was to be expected, performance goes down when only one kind of prosodic stress is considered, until reaching insufficiently reliable values for minor stressed vs. unstressed vocalic nuclei classification. This shows the difficulty of distinguishing between minor stress and the absence of stress and illustrates the difficulty of dealing automatically with many different levels of stress. The discrimination of primary stressed vs. unstressed vocalic nuclei still relies on duration and amplitude for both transcribers' data (Tab. 9). For the S- vs. N classification, average pitch acquires some importance.

Table 8. Primary stressed vs. unstressed (upper rows) and minor stressed vs. unstressed (bottom rows) discrimination: fuzzy logic based algorithm performance. S+ primary, S- minor stressed, N unstressed vowel nuclei.

	data transcr. # 1			data transcr. # 2			average		
	% correct			% correct			% correct		
	S+	S-	N	S+	S-	N	S+	S-	N
TRAINING	100	-	100	100	-	100	100.0	-	100.0
TEST	54	-	88	53	-	83	53.5	-	85.5
TRAINING	-	100	100	-	100	100	-	100.0	100.0
TEST	-	36	85	-	10	95	-	23.0	90.0

5 A Heuristic Algorithm

In the previous sections, we examined the performance of some of the most common data-driven classification methods, like ANNs, statistical decision trees and fuzzy logic based clustering techniques. By applying ANNs to the problem of automatic detection of prosodic stress, the reference performance of a possible algorithm was determined. By applying statistical decision trees and fuzzy logic, the role of the different input features was investigated.

Table 9. Information gain of the input features in the fuzzy models implementing a primary stressed vs. unstressed and minor stressed vs. unstressed vocalic nuclei discrimination on the two transcribers data sets.

	data transcr. # 1				data transcr. # 2			
	information gain for:				information gain for:			
S+ vs. N	duration	amplitude	average pitch	pitch range	duration	amplitude	average pitch	pitch range
1st training set	0.18	0.18	0.05	0.03	0.23	0.11	0.15	0.12
2nd training set	0.34	0.18	0.01	0.02	0.23	0.16	0.06	0.01
3rd training set	0.04	0.15	0.01	0.02	0.21	0.07	0.07	0.11
average	0.19	0.17	0.02	0.02	0.22	0.11	0.09	0.08
	information gain for:				information gain for:			
S- vs. N	duration	amplitude	average pitch	pitch range	duration	amplitude	average pitch	pitch range
1st training set	0.04	0.05	0.01	0.01	0.54	0.29	0.40	0.01
2nd training set	0.12	0.05	0.03	0.03	0.14	0.04	0.01	0.00
3rd training set	0.15	0.04	0.07	0.01	0.09	0.02	0.10	0.00
average	0.10	0.05	0.03	0.01	0.26	0.12	0.17	0.00

In this section, the analysis is performed from a "knowledge" rather than from a data-driven point of view. Let us assume that, to a first approximation, prosodic stress – both primary and secondary – is perceived when a combination of duration, amplitude, average pitch and pitch range of vocalic nuclei produces a local maximum inside the spoken utterance [4]. This could be translated into an automatic algorithm, by defining an evidence variable as a combination of these four acoustic parameters and by detecting its local maxima within the utterance. According to this strategy, after the four acoustic input parameters, described in section 3.2, are expressed in terms of their variance units, they are combined together to form the evidence variable EV_k (Fig. 6).

However, not all local maxima can be considered, because sometimes the speech becomes so soft and almost unintelligible that is not possible to perceive any stress. Thus a local threshold value, T_k, is defined, that states the minimum value of the local maxima to be accepted as pertaining to prosodic stress. Threshold T_k can not be fixed, but has to evolve along with the sentence according to the dynamic of the acoustic features. However, it can not be too flexible, otherwise it follows too closely all possible acoustic variations in the utterance. A compromise is to define threshold T_k as a linear combination of an initial fixed value – T_0 derived from the histogram of the evidence variable over the whole file – and of the average value of the evidence variable over the last n vocalic nuclei, as follows:

$$T_k = a\,T_0 + b\left(\frac{1}{n}\sum_{i=1}^{n} EV_i\right) \tag{11}$$

being $0 \leq a \leq 1$ and $0 \leq b \leq 1$ and $a + b = 1.0$. Parameter a defines the influence of T_0 and parameter b the influence of the average past evidence variable in the definition of threshold, T_k.

Only local maxima EV_k of the evidence variable time series above threshold T_k can be accepted as related to prosodic stress. Let us assume the hypothesis that P% vocalic nuclei are stressed in the utterances. Thus the initial threshold T_0 is determined for each file from the histogram of the evidence variable as this value above which P% of the vocalic nuclei are located (Fig. 7). To verify

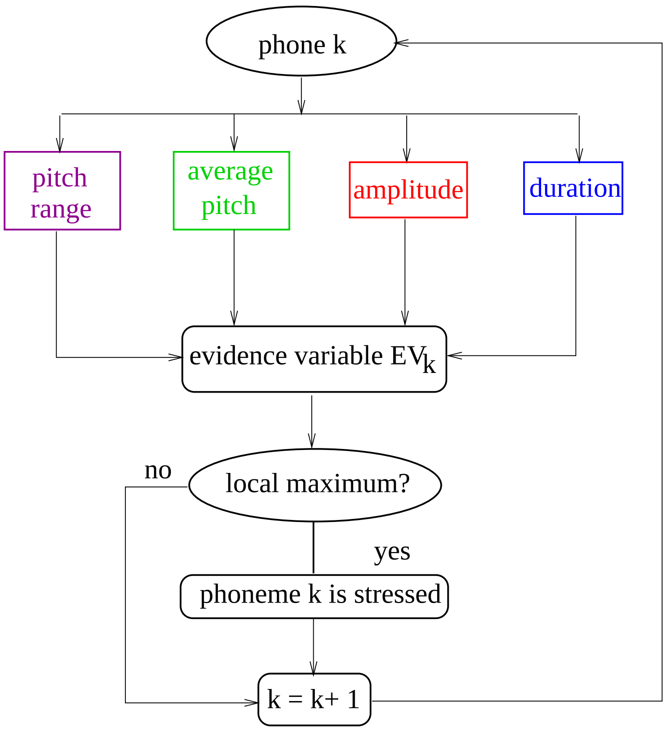

Fig. 6. One possible algorithm for automatic stress detection. Local maxima of the evidence variable, constructed from duration, amplitude, average pitch and pitch range of the vocalic nucleus of every syllable, correspond to stressed syllables.

whether the current value of the evidence variable, EV_k, is a local maximum, it is compared with an α portion of the evidence variable for the previous vocalic nucleus, EV_{k-1}, and with a β portion of the evidence variable for the following vocalic nucleus, EV_{k+1}. EV_k then qualifies as a local maximum if $EV_k \geq \alpha\, EV_{k-1}$ and $EV_k \geq \beta\, EV_{k+1}$. The complete algorithm is summarized in figure 8.

5.1 The training phase

A number of parameters still have to be defined before the algorithm is applied, including the percentage P% of stressed vocalic nuclei in a spoken sentence, the number n of previous vocalic nuclei to keep as reference for the expression of the acoustic features, the portions α and β of the previous and following vocalic nucleus for the local maximum condition, and the coefficients a and b for the threshold updating. A training phase is designed to estimate the optimal values of all these free parameters.

The training is performed separately for each transcriber's dataset on two thirds of the files, in order to assess the best values for the algorithm's parameters. The best values are intended here in terms of the best performance of the algorithm on the training set. The algorithm's performances for the different values of the algorithm's parameters are evaluated by means of Receiver Operator Characteristic (ROC) curves [19].

An ROC curve describes the performance of a system for a two-class discrimination task when

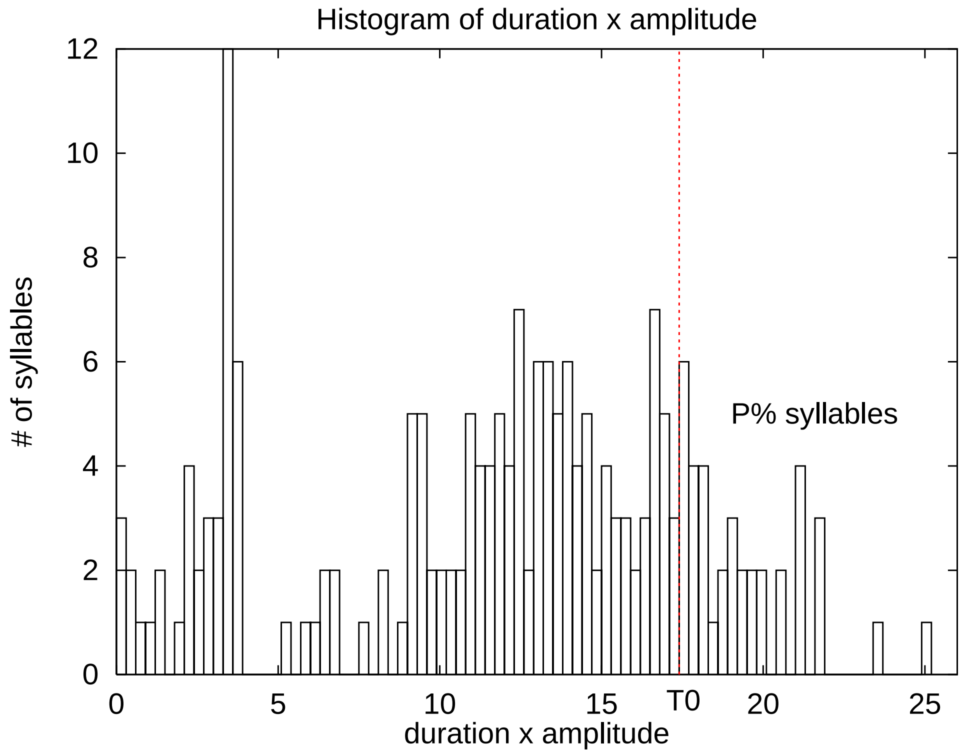

Fig. 7. Histogram of the evidence variable (duration x amplitude) of the entire file shown in figure 1. The threshold T_0 is chosen as that value of the evidence variable above which P% of vocalic nuclei of the sentence are located.

one of the system parameters varies. The proportion of correctly recognized events of each one of the two classes is reported on the x- and y-axis for different values of the varying parameter. A system, correctly classifying every pattern of the two output classes, would have 1.0 on both the x- and y-axes, producing a point on the right upper corner of the graph. In the optimal case, varying one of the system's parameters in one direction causes the proportion of the correctly recognized events of one of the two output classes to decrease, while the other proportion stays constant. Varying the parameter in the other direction yields the opposite effect. Thus, the point on the curve representing the system's performance moves on a line parallel to the x- or to the y-axis respectively. ROC curves are generally used to compare systems' performances. The system with the highest ROC curve produces the best performance.

For the purpose of training, the ROC curves of primary stressed, S+, vs. unstressed nuclei, N, are drawn for different values of n, α, β, a and b. The proportion of S+ vowel nuclei correctly recognized as stressed is reported on the x-axis and the proportion of correctly detected unstressed (N) vowel nuclei on the y-axis while one of the free parameters of the algorithm is varied. The resulting ROC curve gives a measure of the system's performance in classifying primary stressed vowels as stressed (S+) vs. unstressed (N) syllables. The value, producing the point on the curve closest to $[1.0, 1.0]$, is selected for the considered parameter and evidence variable. An example is shown in figure 9, where the evidence variable consists of the product of duration, amplitude and average pitch. The

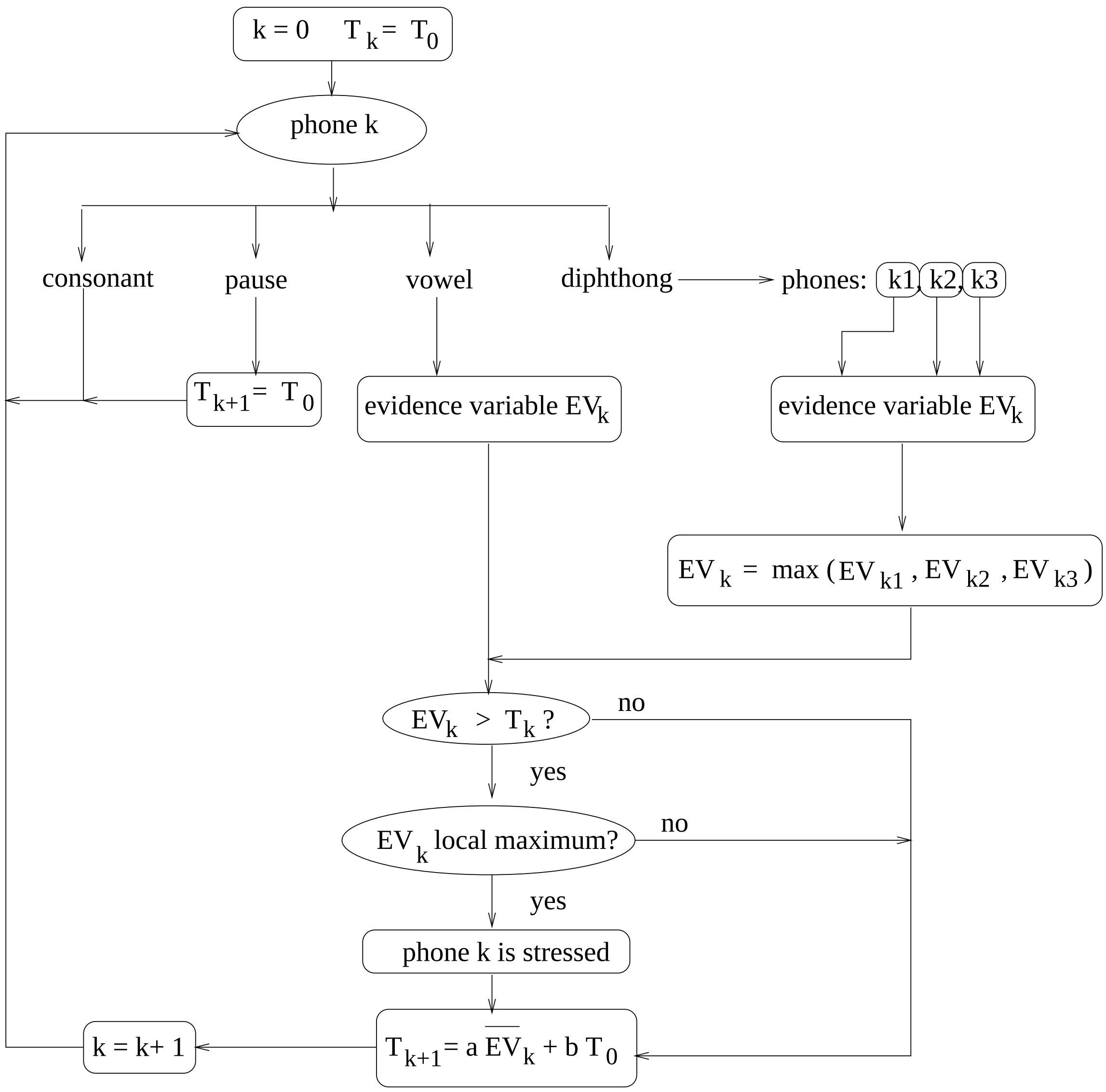

Fig. 8. The complete algorithm proposed for prosodic stress automatic transcription.

free parameters are a and b. The values, $a = 0.5$ and $b = 0.5$, or $a = 0.6$ and $b = 0.4$, yield the point on the ROC curve closest to [1.0, 1.0] and are then chosen for this evidence variable.

For the estimation of the optimal algorithm's parameters, ROC curves are preferred to the most commonly used DET curves for speech recognition [20]. DET curves offer a very good visualization of the error plot only if the error is relatively small (below 30-20%). In the proposed analysis, many of the evaluated evidence variables offer an error around 40-30%, making it hard to visualize of the DET curve.

The test for each evidence variable is finally performed on the remaining one third of the files. In order to give robustness to the algorithm's performance, the Jack-knife procedure is also applied

[14].

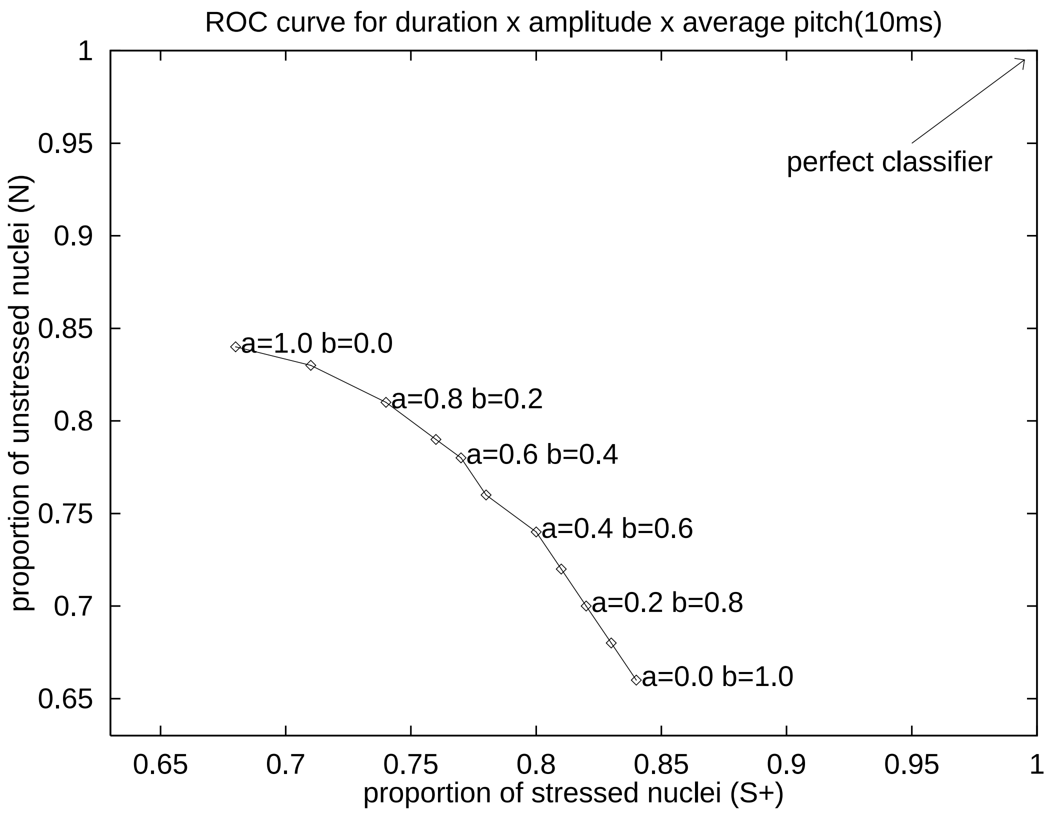

Fig. 9. Example of ROC curve to estimate the best values of a and b taking the product of duration, amplitude, average pitch (10 ms) as the evidence variable. The points with $a = 0.5 b = 0.5$ and $a = 0.6 b = 0.4$ are the closest ones to $[1.0, 1.0]$

As described in section 3.2, two different time windows (25 ms and 15 ms) are used for the estimation of the fundamental frequency $f_0(h)$, leading to 20 ms and 10 ms time grids respectively. The average pitch was calculated initially by using both time grids and the tighter one led to slightly better performance. Hence the 10-ms time grid was used to estimate the fundamental frequency and to calculate the average pitch and pitch-range measures. A number of evidence variables, constructed as a combination of the four acoustic parameters, are evaluated:

- duration (D);
- amplitude (A);
- average pitch with 20-ms time grid (PA(20 ms));
- average pitch with 10-ms time grid (PA(10 ms));
- pitch range (PR);
- pitch range normalized by duration (NPR);
- pitch range x average pitch(10 ms) (PRxPA);
- duration x amplitude;

- duration x pitch range;
- duration x average pitch(10 ms);
- average pitch(10 ms) x amplitude;
- pitch range x amplitude;
- pitch range x amplitude x duration (PRxAxD);
- average pitch(10 ms) x amplitude x duration (PAxAxD);
- normalized pitch range x amplitude x duration (NPRxAxD);
- pitch range x average pitch(10 ms) x amplitude x duration (PRxPAxAxD).

In general, for any adopted evidence variable, the hypothesis of one stressed syllable out of four (P = 25%) leads to the best point on the ROC curve, which confirms what is already known from the literature [4]. Also, the optimum number, n, of past vocalic nuclei to use as reference for the expression of the current value of the evidence variable seems to be quite constant across the different evidence variables and is ca. $n = 15$. Not many different best values are found for α and β, usually between 0.5 and 0.8 for all the evaluated evidence variables. Their value was fixed to 0.6 for both. Parameter a quantifies the contribution of the initial threshold T_0 while parameter b specifies the contribution of the average value of the evidence variable over the past 15 vocalic nuclei in the definition of the adaptive threshold T_k. Both seem to depend on the chosen evidence variable. Table 10 reports the best values for a and b as derived from the ROC curve analysis for the evaluated evidence variables and averaged across the three Jack-knife iterations of training sets.

	data transcr. # 1		data transcr. # 2		average	
evidence variable	a	b	a	b	a	b
duration	0.3	0.7	0.4	0.6	0.35	0.65
amplitude	0.9	0.1	0.8	0.2	0.85	0.15
average pitch (10ms) (PA)	1.0	0.0	1.0	0.0	1.00	0.00
average pitch (20ms)	0.9	0.1	0.9	0.1	0.90	0.10
pitch range (PR)	0.5	0.5	0.4	0.6	0.45	0.55
normalized pitch range (NPR)	0.2	0.8	0.1	0.9	0.15	0.85
PRxPA	0.1	0.9	0.1	0.9	0.10	0.90
duration x amplitude	0.7	0.3	0.7	0.3	0.70	0.30
duration x pitch range	0.0	1.0	0.0	1.0	0.00	1.00
duration x average pitch(10ms)	0.3	0.7	0.3	0.7	0.30	0.70
average pitch(10ms) x amplitude	0.4	0.6	0.4	0.6	0.40	0.60
pitch range x amplitude	1.0	0.0	0.9	0.1	0.95	0.15
PRxAxD	0.1	0.9	0.2	0.8	0.15	0.85
PAxAxD	0.5	0.5	0.6	0.4	0.55	0.45
NPRxAxD	0.1	0.9	0.1	0.9	0.10	0.90
PRxPAxAxD	0.0	1.0	0.1	0.9	0.05	0.95

Table 10. The best a and b parameters for each evidence variable from the ROC curves.

When the four acoustic parameters are used alone, amplitude and average pitch refer to an almost constant threshold $T_k = T_0$ derived from the file's histogram. This means that the amplitude and average pitch of stressed vowels do not depend so much on their past average value. On the other hand, duration and pitch range define the adaptive threshold T_k by taking into account their past average value.

5.2 Training and Test Performance for the Evaluated Evidence Variables

To measure the different discriminative power of the implemented evidence variables, the system's performance on the test and training set were evaluated. In general these two evaluations do not exhibit different percentages, which indicates that no over-training occurs. Moreover, the ROC curves are constructed on the training set by varying the threshold T_k as $q\,T_k$ with $q = 0.0, 0.1, \ldots, 2.0$, and are also compared in order to provide a broader evaluation of the effectiveness of the different evidence variables.

Pitch-Related Acoustic Features. In the literature [4] pitch is cited as one of the most important acoustic features for the perception of prosodic stress by humans. However, extrapolating from human perception to automatic detection may not be warrantal. Indeed pitch variation or high pitch can be a sufficient condition for stress perception, but may not be frequent enough for the automatic recognition of many stressed syllables. In the literature [4] two main pitch-related acoustic features are reported: average pitch and pitch variation (as pitch range or pitch variance) inside a vocalic nucleus. Thus, among the proposed evidence variables average pitch calculated with a time step of 20 ms and 10 ms, pitch range and normalized pitch range are included. Since one of these two groups of features alone might not be sufficiently informative, the product of average pitch and pitch range is also evaluated. In figures 10 and 11, the ROC curves for all the proposed pitch-related acoustic features are depicted for the first and second transcriber's training set, respectively. The corresponding performance on the test sets are reported in table 11.

	transcriber # 1			transcriber # 2			average		
	% correct			% correct			% correct		
	S+	S-	N	S+	S-	N	S+	S-	N
Average Pitch(20ms)	67	57	52	73	61	51	70.0	54.0	51.5
Average Pitch(10ms) (PA)	60	51	60	63	48	60	61.5	49.5	60.0
Pitch Range(PR)	63	53	63	62	50	63	62.5	51.5	63.0
Normalized Pitch Range (NPR)	57	53	57	59	50	60	58.0	51.5	58.5
Pitch Range x Average Pitch (PRxPA)	63	53	67	64	52	64	63.5	52.5	65.5

Table 11. Stressed vs. unstressed discrimination: test set performance of the heuristic algorithm. S+ primary, S- minor stressed, N unstressed vowel nuclei and stress related acoustic features as evidence variables.

Let us compare first the two ways of calculating the average pitch: one with a time grid of 20 ms and the other with a time grid of 10 ms. Not too much can be concluded from the results, since performance associated with the two evidence variables falls into two different ranges for the stressed and the unstressed vocalic nuclei. Some conclusions though can be reached by looking at the corresponding ROC curves (Fig. 10 and 11). For both transcribers' data the improvement of using a 10 ms step instead of a 20 ms step is limited but consistent. Because of that, a 10 ms time step is

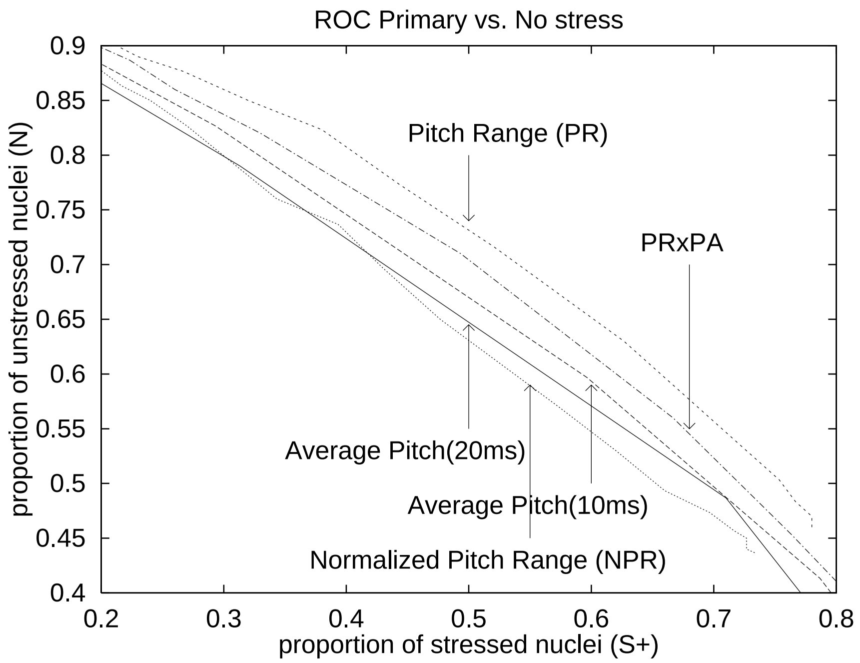

Fig. 10. ROC curves for average pitch, pitch range, normalized pitch range and average pitch x pitch range for a S+ vs. N recognition task (transcriber # 1).

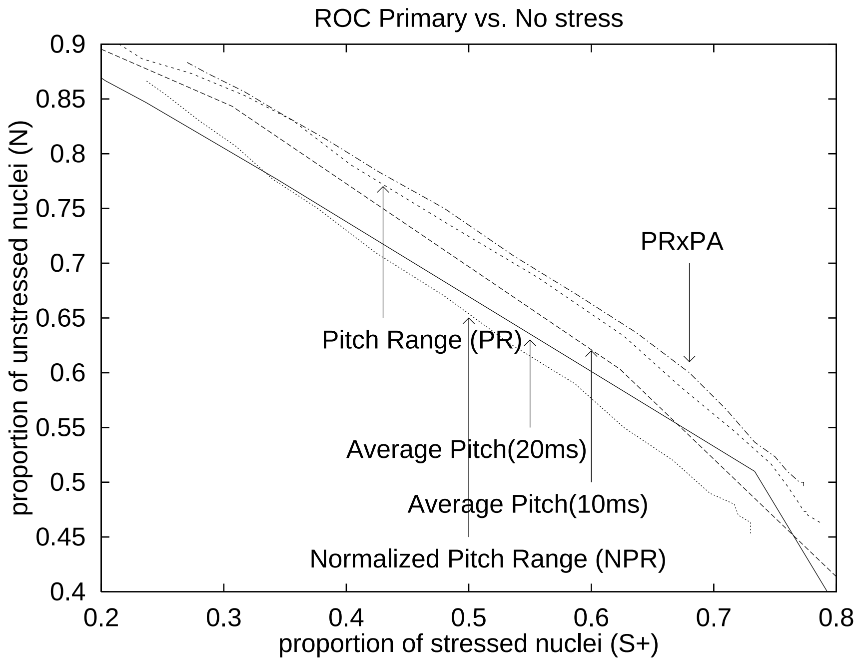

Fig. 11. ROC curves for average pitch, pitch range, normalized pitch range and average pitch x pitch range for a S+ vs. N recognition task (transcriber # 2).

used to estimate the fundamental frequency. Such an improvement, however, disappears when the pitch average is combined with amplitude and/or duration.

The pitch range also offers a constant improvement of some percent in the system's performance with respect to the average pitch for both transcribers' data sets, as it appears from the performance on the test set (Tab. 11) and from the ROC curves (Fig. 10 and 11). This phenomenon appears to be more relevant for the first transcriber's data set than for the second. An objection that can be made is that the higher performance of pitch range is due to the pitch range being related to duration. In fact, a high variation of pitch inside a vocalic nucleus requires a long duration of the vowel. To cope with this objection, the pitch range has been normalized by the duration of the vocalic nucleus. The performance of the normalized pitch range (NMR) decreases dramatically, both on the test set (Tab. 11) and on the training set (Fig. 10 and 11), being the lowest in absolute for all the pitch related features for both transcribers' data sets. Pitch range, without the contribution of duration, does not seem to be a reliable predictor of prosodic stress.

Finally, we have to examine whether pitch range and average pitch carry complementary information. The product of the two offers slightly better performance than average pitch and pitch range alone for the second transcriber's data set, and only intermediate performance between the one of pitch range and the one of average pitch for the first transcriber's data set (Tab. 11 and Fig. 10 and 11). This shows the limited amount of complementary information carried by the two pitch features and the different use of pitch made by the two transcribers to characterize prosodic stress. The first transcriber relies less on pitch information than the second transcriber, granting high information to the pitch range only for the hidden contribution made by duration.

Duration, Amplitude and Pitch. Keeping the two pitch-related acoustic features with the best performance in the previous subsection (that is, pitch range and average pitch with 10-ms time grid) we want to see now how they compare with amplitude and duration (Fig. 12 and 13 and Tab. 12).

	transcriber # 1			transcriber # 2			average		
	% correct			% correct			% correct		
	S+	S-	N	S+	S-	N	S+	S-	N
Average Pitch(10ms)	60	51	60	63	48	60	61.5	49.5	60.0
Pitch Range	63	53	63	62	50	63	62.5	51.5	63.0
Duration	71	60	69	67	56	67	69.0	58.0	68.0
Amplitude	61	51	66	66	47	64	63.5	49.0	65.0

Table 12. Stressed vs. unstressed discrimination: test set performance of the heuristic algorithm. S+ primary, S- minor stressed, N unstressed vowel nuclei and basic acoustic features as evidence variables.

Duration seems to be the best predictor for prosodic stress among the proposed acoustic features for both transcribers and the average pitch is the worst. Amplitude and pitch range are in the middle, amplitude being more important for the second transcriber than for the first one.

In particular, the first transcriber mainly relies on the duration of vocalic nuclei, to recognize primary stress S+. In fact, duration presents in Figure 12 the highest ROC curve on the training set and achieves the best results on the test set (Tab. 12, transcriber # 1) with respect to amplitude,

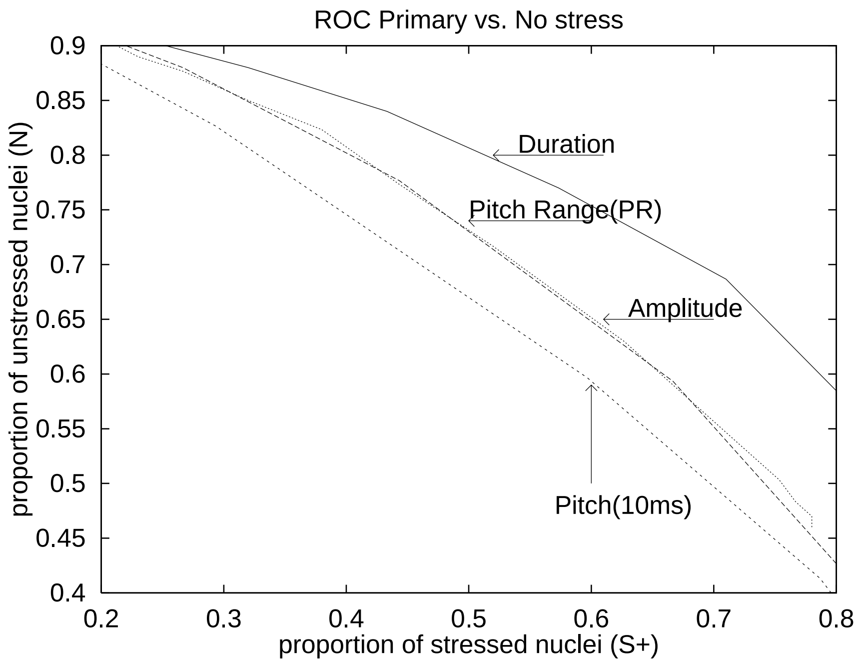

Fig. 12. ROC curves for duration, amplitude, average pitch and pitch range for a S+ vs. N recognition task (transcriber # 1).

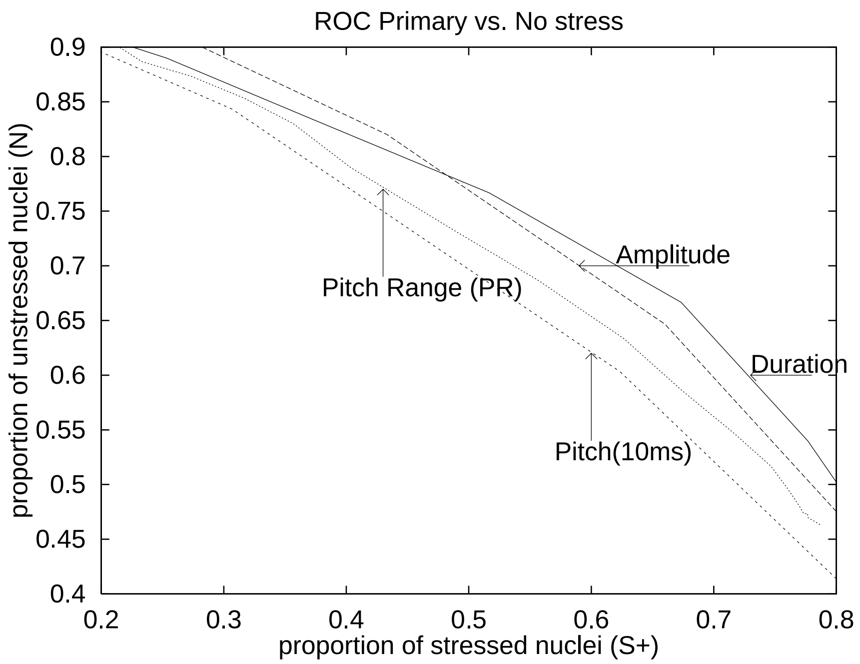

Fig. 13. ROC curves for duration, amplitude, average pitch and pitch range for a S+ vs. N recognition task (transcriber # 2).

average pitch and pitch range. Similar ROC curves to the ones shown in Figure 12, but with lower values of the two proportions, are obtained for the first transcriber using S- and N proportions. For the second transcriber duration is almost as important as the amplitude of the nucleus for primary stress (S+) recognition. This is confirmed by the ROC curves on the training set in figure 13 and by the system's performance on the test set (Tab. 13, transcriber # 2), where duration and amplitude yield the same percentage of correctly recognized S+ vs. N vocalic nuclei. Average pitch also offers better percentages of correctly classified events than for the first transcriber's data set. Duration gains importance in the recognition of intermediate stress (S-).

If the ROC curves and the performance on the test set are evaluated on a subset including only male speakers, duration loses and amplitude and pitch gain some of their discriminative capability in detecting both primary and intermediate stress for both transcribers' data sets.

Products of Pairs. In this subsection the improvement deriving from the combination of two of the basic acoustic features is investigated. Two acoustic features are combined together by means of their product. The ROC curve and the performance on the test set of duration are reported from the previous section, to serve as a reference for the obtained improvement. The ROC curves and the test set performances are reported for Duration x Average Pitch (DxPA), Pitch Range x Duration (PRxD), Duration x Amplitude (DxA), Average Pitch x Amplitude (PAxA), Pitch Range x Amplitude (PRxA) and Pitch Range x Average Pitch (PRxPA) in figures 14 and 15 and in table 13, respectively.

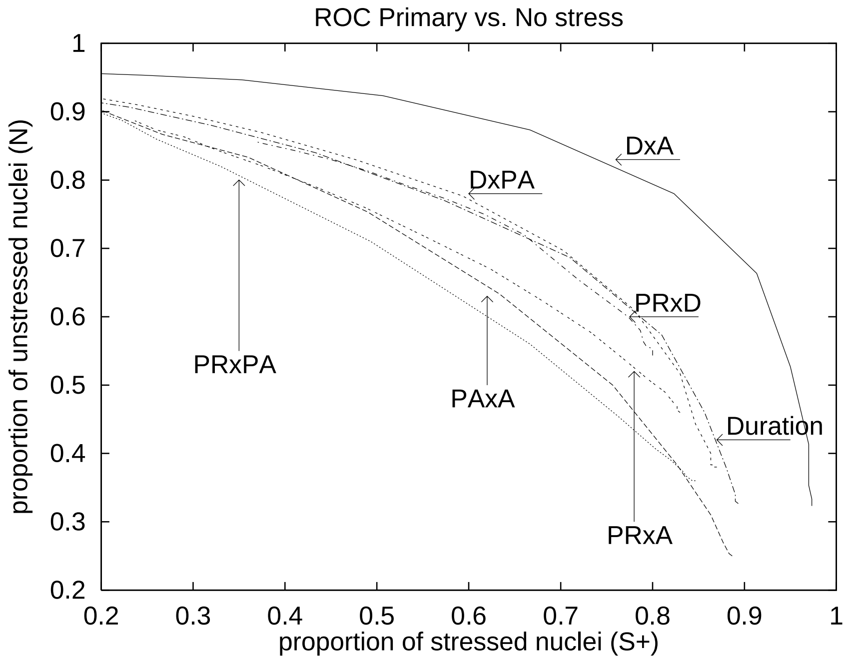

Fig. 14. ROC curves for products of pairs of duration, amplitude, average pitch and pitch range for a S+ vs. N recognition task (transcriber # 1).

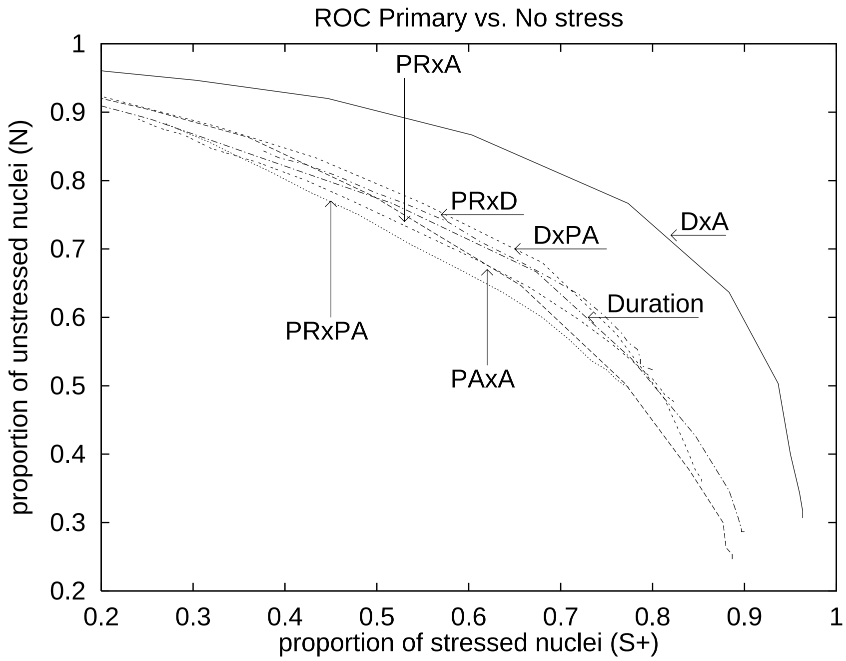

Fig. 15. ROC curves for products of pairs of duration, amplitude, average pitch and pitch range for a S+ vs. N recognition task (transcriber # 2).

Again the product of the two pitch-related features (PRxPA) yields the poorest performance, while the product of duration and amplitude yields the best, both in terms of the ROC curve and in terms of test set performance. In the middle we find the products including amplitude or duration and one of the two pitch-related features.

In addition, the ROC curves of duration x average pitch and amplitude x average pitch have values similar to the ROC curves of duration and amplitude in Figures 12 and 13. This could indicate that average pitch is the least robust or the most redundant acoustic parameter of the vocalic nucleus. Duration x average pitch and pitch range x duration have very close ROC curves as do average pitch x amplitude and pitch range x amplitude, which means that pitch range and average pitch add very similar information to amplitude and duration.

The product of amplitude and duration as an evidence variable dramatically improves the system's performance, yielding 77-81% of correctly identified primary stressed syllables, 59-61% of identified minor stressed and 77-79% of unstressed syllables for the two transcribers' data, respectively (Tab. 13). For the subset of male speakers, all the evidence variables gain a few percent in discrimination capability.

Products of three or more acoustic features. So far, the best performance is obtained by using the product of duration and amplitude. Could this performance be improved, by introducing one or more acoustic features in this product? In order to test this hypothesis, the performance of average pitch x amplitude x duration (PAxAxD), pitch range x amplitude x duration (PRxAxD) and pitch range x average pitch x amplitude x duration (PRxPAxAxD) are compared with the performance of duration x amplitude as well as of duration alone (Fig. 16 and 17, Tab. 14).

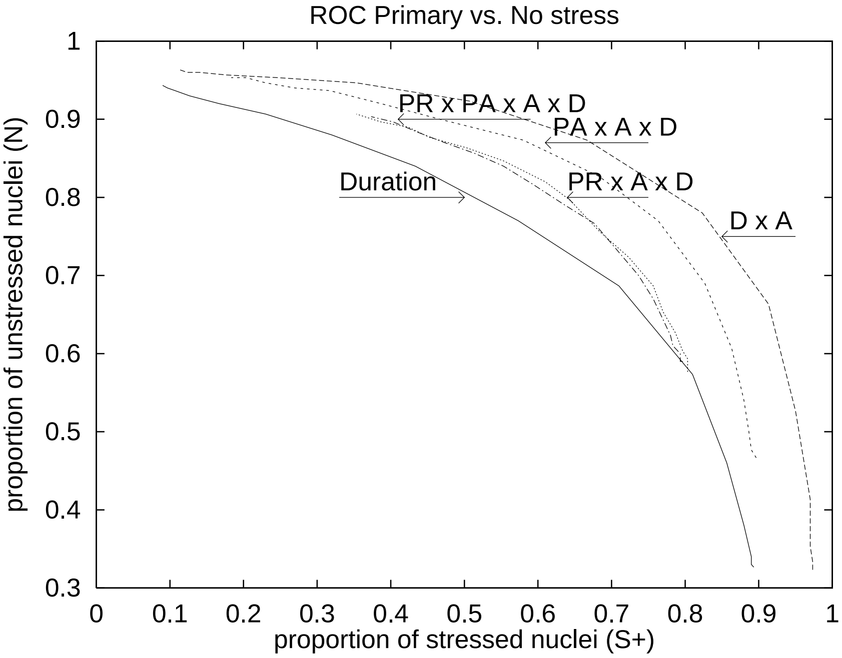

Fig. 16. ROC curves for products of three or more features among duration, amplitude, average pitch and pitch range for a S+ vs. N recognition task (transcriber # 1).

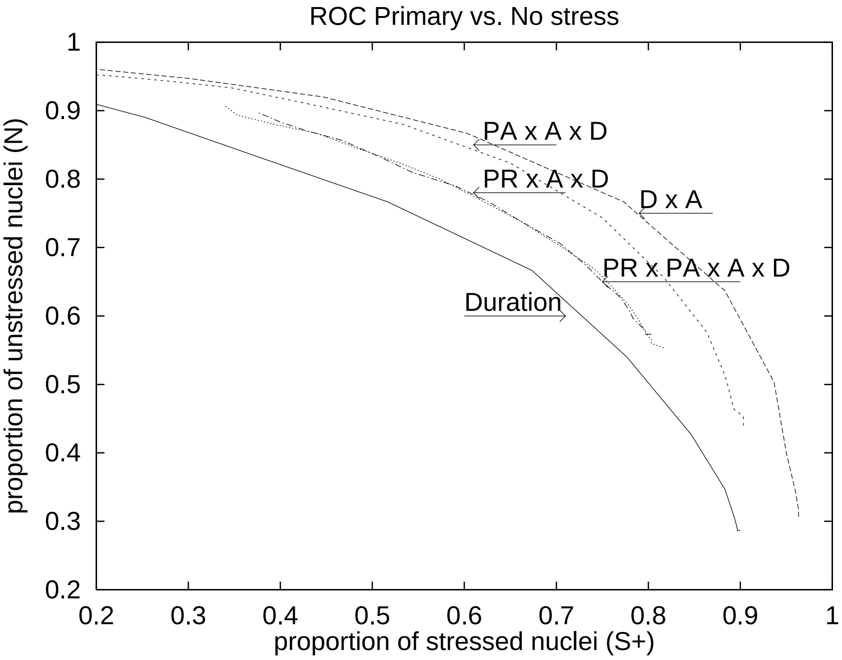

Fig. 17. ROC curves for products of three or more features among duration, amplitude, average pitch and pitch range for a S+ vs. N recognition task (transcriber # 2).

	transcriber # 1			transcriber # 2			average		
	% correct			% correct			% correct		
	S+	S-	N	S+	S-	N	S+	S-	N
Duration x Amplitude (DxA)	81	61	79	77	59	77	79.0	60.0	78.0
Duration x Average Pitch (DxPA)	71	58	70	67	57	68	69.0	57.5	69.0
Pitch Range x Duration (PRxD)	70	59	67	69	59	65	69.5	59.0	66.0
Amplitude x Average Pitch (AxPA)	63	50	63	66	49	65	64.5	49.5	64.0
Pitch Range x Amplitude (PRxA)	69	58	62	68	53	63	68.5	55.5	62.5
Pitch Range x Average Pitch (PRxPA)	63	53	67	64	52	64	63.5	52.5	65.5
Duration	71	60	69	67	56	67	69.0	58.0	68.0

Table 13. Stressed vs. unstressed discrimination: test set performance of the heuristic algorithm. S+ primary, S- intermediate stressed, N unstressed nuclei and products of a pair of basic acoustic features as evidence variables.

	transcriber # 1			transcriber # 2			average		
	% correct			% correct			% correct		
	S+	S-	N	S+	S-	N	S+	S-	N
Duration	71	60	69	67	56	67	69.0	58.0	68.0
DxA	81	61	79	77	59	77	79.0	60.0	78.0
PAxAxD	76	56	75	75	56	75	75.5	56.0	75.0
PRxAxD	72	58	75	70	56	70	71.0	57.0	72.5
PRxPAxAxD	71	55	73	70	55	71	70.5	55.0	70.5

Table 14. Stressed vs. unstressed discrimination: test set performance. S+ primary, S- intermediate stressed, N unstressed nuclei and products of three or more basic acoustic features as evidence variables.

Duration x amplitude still produces the best performance. The introduction of average pitch and even more the introduction of average pitch x pitch range, reduces the percentage of correctly classified events slightly but consistently (Fig. 16 and 17) for both transcribers' data sets.

From the results in Table 13, the vocalic nuclei seem to contain sufficient information, in terms of duration, amplitude and pitch, for a good discrimination of S+ and N syllables, both around 80% for both transcribers' data. Intermediate stresses S- are less reliably detected (59-61%) on the basis of vocalic information alone.

The system's performance on the training set are always very close to the system performance on the test set. Indeed, the small number of the algorithm's free parameters does not allow over-fitting of the training data, granting generality to the conclusions derived from the ROC curves about the role of pitch features, amplitude and duration in prosodic stress recognition. The introduction of additional free parameters in the simple structure of the algorithm in figure 6 could allow the implementation of better discrimination surfaces, resulting in an improvement of the system's performances.

6 Summary

An automatic algorithm for marking prosodic stress in spontaneous American English discourse was designed, using different data-driven and knowledge-based analysis techniques. The analysis of some of these techniques allow the investigation of prosodic-stress properties of syllabic sequences, in terms of duration, amplitude and pitch. The evaluation is performed on two separate subsets of the OGI Corpus, partially overlapping, and separately labeled by two transcribers.

All the interpreted data-driven and knowledge-based techniques lead to the same conclusion. The duration of the vocalic nuclei seems to play a major role in prosodic stress characterization, followed in order of importance by amplitude, average pitch and normalized pitch range. Pitch range alone apparently performs better than average pitch, because of its correlation with duration and amplitude. The best performance is obtained by using the product of duration and amplitude as an evidence variable and is only slightly worse than the agreement percentages between the two transcribers.

7 Acknowledgments

This research was supported by the US Dept of Defense. We thank Jeff Good and Joy Hollenback for the hand-labeled stress annotations used in this study.

References

1. Lehiste I. 1970. *Suprasegmentals* MIT Press, Cambridge.
2. Kuijk, D. van and Boves, L. 1999. Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication* 27, 95-111.
3. Wightman, C.W. and Ostendorf, M. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing* 2,469-81.
4. J. Clark and C. Yallup, "Introduction to Phonology and Phonetics", Oxford: Blackwell, 1990.
5. Bergem, D. van 1993. Acoustic vowel reduction as a function of sentence accent, word stress and word class on the quality of vowels. *Speech Communication* 12, 1-23.
6. R. Silipo, "Artificial Neural Networks", in: Intelligent Data Analysis: An Introduction, M. Berthold and D. Hand eds, Springer-Verlag, Berlin-Heidelberg, pp. 219-272, 1999.
7. J.R. Quinlan, "Induction of Decision Trees", in *Machine Learning*, pp. 81-106, 1986.
8. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.
9. L.A. Zadeh, "A fuzzy-algorithmic approach to the definition of complex or imprecise concepts", *Int. J. Man-Machine Studies*, **8:** 249-291, 1976.
10. R. Silipo and S. Greenberg, "Automatic transcription of prosodic stress for spontaneous English discourse", Proc. of the XIVth International Congress of Phonetic Sciences (ICPhS), 3:2351, 1999.
11. Center for Spoken Language Understanding, Dept. of Computer Science and Engineering, Oregon Graduate Institute. 1995. *Stories corpus*, Release 1.0.
12. N. Chomsky and M. Halle, *The Sound Pattern of English.* New York: Harper and Row, 1968.
13. Hess, W. 1983. *Pitch Determination of Speech Signals: Algorithms and Devices.* Berlin, Springer-Verlag.
14. A.J. Feelders, "Statistical Concepts", in: *Intelligent Data Analysis: An Introduction*, M. Berthold and D. Hand eds, Springer-Verlag, Berlin-Heidelberg, pp. 15-66, 1999.
15. C.Z. Janikow, "Fuzzy Decision Trees: Issues and Methods", *IEEE Trans. Syst. Man and Cyb. Part B: Cybernetics*, **28:** 1-14, 1998.
16. K.-P. Huber and M.R. Berthold, "Building Precise Classifiers with Automatic Rule Extraction", in Proceedings of the *IEEE International Conference on Neural Networks*, vol. 3, pp. 1263-1268, 1995.

17. M. R. Berthold, K.P. Huber, “Comparing Fuzzy Graphs”, Proc. of Fuzzy-Neuro Systems, pp. 234-240, 1998.
18. R. Silipo and M. Berthold, ”Discriminative Power of Input Features in a Fuzzy Model”, *LNCS1642*, Springer-Verlag, pp. 85-96, 1999.
19. Green, D. M. and Swets, J.A. 1966. *Signal detection theory and psychophysics.* New York, Wiley.
20. NIST Spoken Language Technology Evaluation and Utility Software, http://www.itl.nist.gov/iaui/894.01/software.htm