AUTOMATIC TRANSCRIPTION OF PROSODIC STRESS FOR SPONTANEOUS ENGLISH DISCOURSE

Rosaria Silipo Steven Greenberg
International Computer Science Institute
1947 Center Street, Berkeley CA 94704, USA
email: rosaria,steveng@icsi.berkeley.edu

ABSTRACT

The role of duration, amplitude and fundamental frequency of syllabic vocalic nuclei is investigated for marking prosodic stress in spontaneous American English discourse. Local maxima of different evidence variables, implemented as combinations of the three basic parameters – duration, amplitude and pitch –, are supposed to be related with prosodic stress. As reference, two different subsets from the OGI English stories database were manually marked in terms of prosodic stress by two different trained linguists. The ROC curves, built on the training examples, show that both transcribers grant a major role to the amplitude and duration rather than to the pitch of the vocalic nuclei. More complex evidence variables, involving a product of the three basic parameters, allow around 80% primary stressed and 77% unstressed syllables to be correctly recognized in the test files of both transcribers' datasets. The agreement between the two transcribers on a set of common files supplies only slightly higher percentages.

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. 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.

These three parameters assume very different values across the consonant 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 large part of prosodic stress information is carried by the vocalic nucleus [4, 2] and in order to avoid complicated normalization problems, the role of duration, amplitude and fundamental frequency of solely syllabic vocalic nuclei 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. Different combinations of these three basic parameters of the vocalic nuclei are also evaluated, to assess their effectiveness and reliability for prosodic stress detection.

2. MARKING PROSODIC STRESS

2.1. Amplitude, duration and pitch

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.

Inside a speech file, the *duration* of the k-th vocalic nucleus is the number, D_k, of signal samples between its onset and end.

The *amplitude*, A_k, is defined as the Root Mean Square of the D_k signal samples contained in the k-th vocalic nucleus.

Finally, the *pitch*, P_k, refers to the average value of the fundamental frequency, $f_0(t)$, inside the k-th vocalic nucleus. Fundamental frequencies $f_0(t)$ are estimated on the basis of the autocorrelation function of quarter of octave spectral channels, calculated on a 25 ms time window centered around time t and overlapping 5 ms with the previous and following time windows [6]. If N_k such fundamental frequencies, corresponding to N_k partially overlapping 25 ms time windows, are detected inside the k-th vocalic nucleus, the corresponding pitch P_k is evaluated as to their average value (eq. 1). This technique neutralizes residual outliers reflecting transitions from vowel to consonant and viceversa.

$$P_k = \frac{1}{N_k} \sum_{t=1}^{N_k} f_0(t) \tag{1}$$

2.2. The stress assignment procedure

For the stress assignment procedure, only two classes of 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", "oy", "er", present a longer duration than plain vowels and, because of that, are divided in three parts. For the same reason, artificially elongated vowels, that is longer than 25 ms or 40 ms, are split into three and five parts respectively. The maximum value, assumed by the evidence variable over all the resulting parts, is retained for the analysis.

Every speaker appears to use different combinations of duration, amplitude and pitch, to produce stressed vowels. In order to normalize this variance among speakers, duration, amplitude and pitch are expressed in terms of variance units from the mean value of their probabilistic distributions inside each utterance.

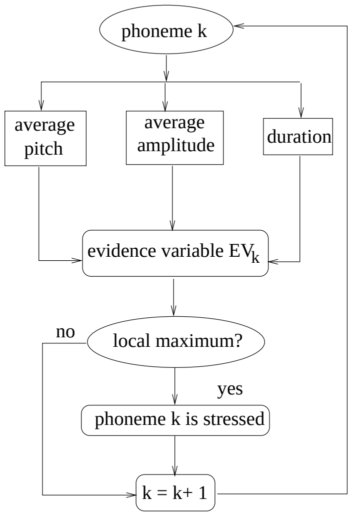

Figure 1. The algorithm for automatic stress detection. Local maxima of the evidence variable, built from duration, amplitude and pitch of the vocalic nucleus of every syllable, correspond to stressed syllables.

Based on the concurrent action of pitch, duration and amplitude of each syllabic vowel nucleus, an evidence variable, EV_k, is defined. To a first approximation, local maxima in the evidence variable time series correspond to stressed vowels inside a sentence.

During the training phase, for each file a threshold value, T_0, is extracted from the histogram of the evidence variable, as that value for which $EV_k < T_0$ holds for a certain proportion P of the file's vowels.

During the test phase, T_0 is the initial recognition threshold. Later on, an adaptive threshold, T_k, is used. T_k is defined as a linear combination of the histogram threshold, T_0, and the average value of the evidence variable in the past 15 vocalic nuclei. If the evidence variable, EV_k, of the k-th vocalic nucleus is above threshold T_k, vowel k is a good candidate for carrying stress. The detection threshold T_k is set again to T_0 after every pause in the sentence.

The second step in this process consists of verifying whether the corresponding evidence variable is a local maximum. The evidence variable, EV_k, must be above the α portion of the evidence variable of the previous vocalic nucleus ($EV_k > \alpha\ EV_{k-1}$), above the β portion of the evidence variable of the following vocalic nucleus ($EV_k > \beta\ EV_{k+1}$), and above the γ portion of the average value of the evidence variable in the previous and the following vocalic nucleus ($EV_k > \gamma\ \frac{EV_{k-1}+EV_{k+1}}{2}$).

The following vowel $k+1$ is then examined and the procedure is repeated. The algorithm is summarized in Figure 1.

3. THE TRANSCRIBED CORPUS

To provide a reference platform for the algorithm's performance, the prosodic stress of a portion of the American English component of the OGI Stories Corpus [7] was manually marked by two trained linguists.

The corpus contains 50-60 second files about any subject. A phonetic transcription of the files was also supplied. Two different subsets of files were 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 52 files, with 39 men and 13 women voices. 10 files, 5 men and 5 women voices, represent the overlapping part of the two subsets. The evaluation of the algorithm was independently performed on the two data subsets.

The annotations refer both to primary stressed (S+), to other minor stressed (S-), and to unstressed syllables (N). The automatic classification described in the previous section focuses on the recognition of primary stress (S+) vs. unstressed syllables.

4. PERFORMANCES OF DIFFERENT EVIDENCE VARIABLES

4.1. The ROC curves

The automatic stress detection algorithm was evaluated by using the single parameters (duration, amplitude, pitch), their paired products, two by two, and the product of all of them together. The corresponding performance is described by means of the Receiving Operator Characteristic (ROC) curve [5].

The ROC curve produces a measure of the system's performance, when one of its parameters varies. In a two-class discrimination task, the proportion of correctly recognized events of one of the two classes is reported on the x-axis and the proportion of the correctly recognized events of the other class is reported on the y-axis. A system, correctly classifying every pattern of the two classes, has 1.0 both on the x- and y-axis, producing a point on the right upper corner of the graphic. 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 classes to decrease, while the other proportion stays constant. Varying the parameter on 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 curve produces the best performance.

In the stressed (S) vs. unstressed (N) discrimination task, performed by the proposed stress detection algorithm, the proportion of vowel nuclei S+ detected by the algorithm as S is reported on the x-axis and the proportion of correctly detected unstressed (N) vowel nuclei on the y-axis. The resulting ROC curve gives a measure of the system's performance in classifying primary stressed vowels as stressed (S) vs. unstressed (N) vowels. If the focus of the analysis is on minor stresses, the proportion of vowel nuclei S- labeled as S by the algorithm will measure the algorithm's capability in detecting minor stressed vowels as stressed (S) vs. unstressed (N) vowels.

4.2. The training phase

The training is performed separately for each transcriber's dataset on two thirds of the files, in order to determine: 1) the best proportion, P, of unstressed vowels in each file, for the definition of the initial threshold T_0; 2) the most appropriate value for the coefficients, a and b, involved in the calculation of the threshold T_k. b defines the contribution of the 15 previous evidence variable values, while a is the contribution of the initial threshold, T_0.

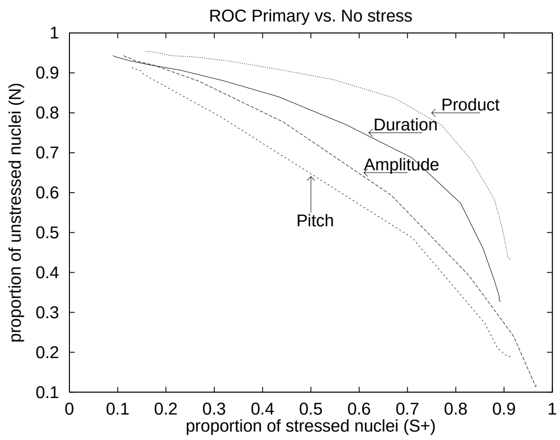

Figure 2. ROC curves for duration, average amplitude, and average fundamental frequency for a S+ vs. N recognition task (transcriber # 1).

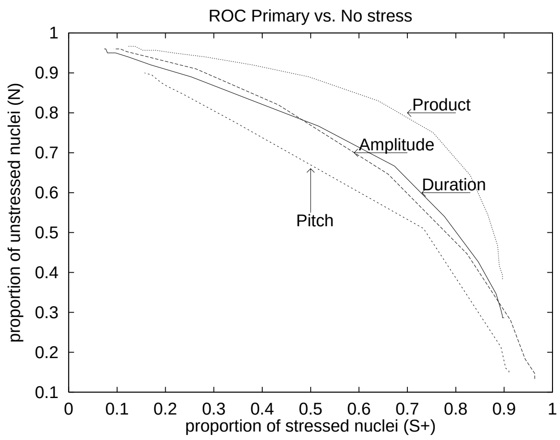

Figure 3. ROC curves for duration, average amplitude, and average fundamental frequency for a S+ vs. N recognition task (transcriber # 2).

The ROC curves of primary stressed S+ vs. unstressed vowels N are drawn for different values of P, a and b. The values, producing the closest point to $[1.0, 1.0]$, are then selected. Usually, for any adopted evidence variable, the hypothesis of one stressed syllable out of four ($P = 0.75$) leads to the best point on the ROC curve. The remaining one third of the data subset is used as test set. Different training and test sets were evaluated, by using the Jackknife method. The evaluation results on the test set are reported in Table 1 for different evidence variables.

4.3. Duration, amplitude, and fundamental frequency

To measure the different discriminative power of the implemented evidence variables, after the training procedure is ended a ROC curve is built on the training data, by varying the threshold T_k as $q\ T_k$ with $q = 0.0, 0.1, \ldots, 2.0$. In Figures 2 and 3, the ROC curves for duration, amplitude and pitch are depicted for the first and second transcriber's training set respectively. The ROC curve of the product of the three parameters is also reported as reference.

The first transcriber mainly relies on duration of vowel nuclei, to recognize primary stress S+. In fact duration presents in Figure 2 the highest ROC curve on the training set and achieves the best results on the test set (Tab. 1, transcriber # 1) with respect to amplitude and pitch. Similar curves to the ones in Figure 2, but with lower values of the two proportions, are obtained for the first transcriber considering S- and N proportions.

For the second transcriber the duration is as important as the amplitude of the vowel nucleus for primary stress (S+) recognition. This is confirmed by the ROC curves on the training set in Figure 3 and by the system's performance on the test set (Tab. 1, transcriber # 2), where duration and amplitude yield the same percentages of correctly recognized S+ vs. N. Duration gains again importance in the recognition of minor stresses (S-).

The worst performances for both transcribers is in terms of pitch, both as ROC curves on the training set (Fig. 2 and 3) and as performances on the test set (Tab. 1).

If the ROC curves and the performance on the test set are evaluated on a more homogeneous subset including only male voice speaker, duration loses and amplitude and pitch gain part of their discriminative capability, in detecting both primary and minor stresses.

4.4. The products

The ROC curves of duration x pitch and amplitude x pitch show the same trend as the ROC curves for duration and amplitude in Figures 2 and 3. This could indicate that f_0 is the least robust or the most redundant parameter of the vocalic nucleus. Consistent with this hypothesis, the product of the three single parameters produces a lower or a comparable ROC curve to the one pertaining to amplitude x duration. Moreover the algorithm's performance on the test set (Tab. 1) show lower percentages for the product of the three basic parameters than for duration x amplitude.

The product of amplitude and duration as an evidence variable dramatically improves the system's performance, yielding a 81-77% of correctly identified primary stressed syllables, a 61-59% of identified minor stressed and a 79-77% of unstressed syllables for the two transcribers' data respectively (Tab. 1).

From the results in Table 1, the vocalic nuclei seem to contain sufficient information, in terms of duration, amplitude and f_0, for a good discrimination of S+ and N syllables, both around 80% for both transcribers' data. Minor stresses S- are less reliably detected (61-59%) on the basis of the solely vocalic information. The best results on the test set are obtained by using the product of amplitude and duration as evidence variable.

For a subset of male only speakers, all the evidence variables gain a few percent in discrimination capability.

The algorithm's performance on the training set are very similar to the ones on the test set (Tab. 1). The low number of algorithm's free parameters does not allow any overfitting of the training data, granting generality to the conclusions derived from the ROC curves about the role of pitch, amplitude and duration in prosodic stress recognition. The introduction of new free parameters in the simple structure of the algorithm in Figure 1 could allow the implementation of better discrimination surfaces and then an improvement of the final performances.

	transcriber # 1			transcriber # 2		
	% correct			% correct		
	S+	S-	N	S+	S-	N
Product	77	57	77	75	58	75
Dur. x Amp.	81	61	79	77	59	77
Dur. x Pitch	71	58	70	67	57	68
Amp. x Pitch	63	50	63	66	49	65
Duration	71	60	69	67	56	67
Amplitude	61	51	66	66	47	64
Pitch	67	57	52	73	61	51

Table 1. Stressed vs. unstressed discrimination: test set performances. S+ primary, S- minor stressed, N unstressed vowel nuclei.

	Transcr. # 1 vs. # 2			Transcr. # 2 vs. # 1		
	% correct			% correct		
	S+	S-	N	S+	S-	N
W+M	90	67	84	78	57	93
M	93	76	84	81	58	94
W	87	46	85	74	56	92

Table 2. In the first three columns, agreement of transcriber # 1 vs. transcriber # 2 and, in the last three columns, agreement of transcriber # 2 vs. transcriber # 1 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.

4.5. Transcribers' agreement

The agreement between the two transcribers on the common files of the two OGI data subsets is shown in Table 2, to compare the algorithm's performance with the transcribers' agreement.

The first three columns of Table 2 refer to the agreement percentage of transcriber # 1 vs. transcriber # 2, considering only men voice files (M), only women voice files (W) and both together (W+M). The second three columns refer to the agreement percentage of transcriber # 2 vs. transcriber # 1. In order to compare the agreement with the algorithm's results, a 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 two transcribers roughly agree in recognizing unstressed syllables (N: 84-93%) and primary stress (S+: 90-78%). Much more disagreement exists in recognizing minor stresses (S-: 67-57%). Many syllables marked by transcriber # 1 as minor stressed are labeled by transcriber # 2 as S+ stresses. In general, transcriber # 2 seems to be more biased towards marking primary stress than transcriber # 1. The strongest disagreement about S- stresses regards female speakers.

The algorithm's performance is encouraging, if compared with the agreement percentages of the two transcribers in a primary stressed vs. unstressed syllables discrimination task. Finer discriminations among different kinds of stress do not yield yet a sufficient amount of agreement among human transcribers, to encourage an automatic implementation.

5. DISCUSSION

Other linear and nonlinear combinations of amplitude, duration and pitch were evaluated on a subset of the Switchboard Corpus, including the 23 longest files of the Switchboard database, during a preliminary phase of the experiments. Despite the heavier computation required, such more elaborated evidence variables showed worse or comparable performances to the ones of duration x amplitude, confirming in general the secondary role of pitch and the major role of duration for prosodic stress recognition.

An attempt to discriminate S+ from S- stresses, inside the group S of stressed detected syllables, was implemented on the same subset of the Switchboard database. The histogram of primary and minor stressed vowels does not show any promising chances, to define a robust and reliable threshold for S+ vs. S- stress discrimination. In this first attempt, a fixed threshold, T_s, was defined file by file, based on the three highest values of the evidence variable in each file. If a vowel, already labeled as stressed S, shows an evidence variable higher than T_s, its stress is classified as S+, otherwise as S-. For this S+ vs. S- stress discrimination task, very poor results were obtained. However the best performances came from evidence variables incorporating pitch, as for example 80% vs. 57% for the pitch alone.

6. SUMMARY AND CONCLUSIONS

An automatic algorithm for marking prosodic stress in spontaneous American English discourse investigates the prosodic stress properties of vocalic nuclei 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 trained linguists.

The duration of the vocalic nuclei seems to play a major role in prosodic stress characterization, followed in importance by amplitude and pitch. The best performances are obtained by using the product of duration and amplitude as evidence variable and are only slightly worse than the agreement percentages between the two transcribers.

ACKNOWLEDGMENTS

This research was supported by the 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] 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.

[5] Green, D. M. and Swets, J.A. 1966. *Signal detection theory and psychophysics.* New York, Wiley.

[6] Hess, W. 1983. *Pitch determination of speech signals: algorithms and devices.* Berlin, Springer-Verlag.

[7] Center for Spoken Language Understanding, Dept. of Computer Science and Engineering, Oregon Graduate Institute. 1995. *Stories corpus*, Release 1.0.