Recognition of Lexical Stress in a Continuous Speech Understanding System - A Pattern Recognition Approach

Alex Waibel

Computer Science Department
Carnegie-Mellon University
Pittsburgh, PA 15213

Abstract

Stress is one of the key components in human speech perception. Its uses extend from the phonetic level over the lexical to the syntactic and semantic level. Several methods have been developed in the past to detect stress automatically from the signal. This paper takes a pattern recognition approach to the the problem of stress detection. The algorithm presented has three key features: (1) optimal combination of the evidence obtained from the acoustic correlates of stress is achieved by means of a Bayesian classifier assuming multivariate Gaussian distributions; (2) the algorithm detects lexical stress in continuously spoken English utterances; (3) rather than making hard decisions, the algorithm returns probabilities for each syllable, i.e., a measure of stressedness. The algorithm was tested over 4 databases of differing continuous speech data. When a forced decision is imposed by setting a threshold at stress probability 0.5, error rates of 7.79% to 14.85% missed stresses were obtained. Unlike in other languages (such as Japanese), amplitude integrals are the strongest predictor of English stress. Performance results and an analysis of errors are presented.

1. Introduction

Stress has repeatedly been found to be an extremely important factor in speech perception. Stressed syllables are usually the best articulated syllable in a word and thus could provide islands of phonetic reliability [1, 2]. With a decrease in degree of stressedness (e.g., reduced syllables at higher speaking rates) all vowels appear to move towards a neutral, central schwa-like point in F1/F2-space or are deleted altogether. Stressed syllables in a large English dictionary also carry more acoustically discriminatory information than unstressed syllables [3] and therefore provide not only acoustic reliability, but also more discriminatory information content. Moreover, in sentence context, content words, i.e., "important" words carrying most of the semantic information content of a sentence, are mostly stressed, while function words (articles, conjunctions, possessive determiners, etc.) tend to be unstressed or reduced. In English, word-level stress is free, i.e., its position is not fixed within the word. Stress could thus also be used as a constraint for lexical access. This latter property has in fact been demonstrated to be of potential value for word hypothesization in automatic large vocabulary recognition [4, 5]. The analysis of rhythmic and/or temporal aspects of speech also relies heavily on the availability of stress information. Speaking rate and the corresponding applicability of phonological variations such as palatalization, glottal stop and pause insertion, flapping, etc. can be best predicted by the intervals between stressed syllables, the so-called interstress intervals [1]. Since English is said to be a stress-timed language, these interstress intervals are approximately isochronous and their duration determines the rate of speech. Significant deviations from these interstress intervals typically indicate major syntactic boundaries such as phrase boundaries or clause boundaries [1].

This research was sponsored in part by the National Science Foundation, Grant MCS-7825824 and in part by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-78-C-1551.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

2. Acoustic Correlates of Stress

Although the importance of the percept of stress in speech perception and speech recognition seems intuitively clear from the foregoing discussion, the acoustic manifestations of stress have been the subject of debate and appear to be language dependent. Various perceptual studies have found pitch to be the most salient correlate of stress in English. Conversely, measurements of acoustic features in word pairs that differ only in stress pattern (CONvict/conVICT) indicate that the amplitude integral is the strongest correlate of stress. (For a good review see Lehiste [6].) Nevertheless, in contrast to other languages such as Japanese (where only pitch accent is systematically used [7]), English has been found to use all of the above acoustic features as correlates of stress.

In the experiments described in the following, these predictions and observations will be reexamined to arrive at an optimal automatic stress detection algorithm. Rather than relying on manually set thresholds, we will then use an automatic learning algorithm to obtain an optimal combination of these acoustic correlates for automatic stress detection. Results of extensive evaluation of over 2850 syllables from continuously spoken sentences will also be given below.

3. Automatic Detection of Stress

The notion of stress or prominence is so intuitive that this frequently clouds the fact that automatic stress detection itself is a difficult problem. Several attempts have been made to automatically detect stress in spoken English. Respectable performance was achieved for isolated word stress detection by Lieberman in 1960 [8] and Aull [5]. Lea [1] reports good performance for a continuous speech stress detection algorithm. In the following sections we will attempt to construct a stress detection algorithm that operates on continuous speech, attempts to detect lexical stress, and assigns a probability of stressedness to all syllables. A pattern recognition approach is adopted to optimally combine various features (parameters acting as correlates of stress) into one minimum-error stressed-syllable classifier. Since this

CH2243-4/86/0000-2287 $1.00

approach departs in various ways from previous systems we will start by first reexamining the acoustic correlates of stress found in our database of continuously spoken sentences. These are intensity, duration, pitch, and spectral change. For this analysis as well as for the final stress detector a database of 50 sentences (Harvard sentences) read by 5 speakers (3 male, 2 female, reading 10 sentences each) was used. Each sentence was labeled according to three coarse phonetic classes: silence, fricative and vocalic. In addition, syllable boundaries and stress labels were added to these label files. All stress labels were *lexical* stress labels as indicated by a dictionary or as derived synthetically by rule. Only 3 levels of stress were used: primary stress, secondary stress, and unstressed. This labeling scheme, of course, ignores sentential stress, emphasis, phrase level phenomena, rhythmic/syntactic/semantic phenomena, that all do indeed play a significant role on the actual realization and perception of stress in spoken English sentences. In the following, each of the 50 utterances was automatically segmented into syllables. For each syllable a particular feature was measured and pooled into the stressed or unstressed category in order to compute histograms and class conditional probability density functions of a given feature for stressed and unstressed classes (secondary stresses were considered to be stressed). In this fashion a training database of 244 stressed syllables and 238 unstressed syllables was obtained. To achieve nearly Gaussian distributions, each feature is offset by a constant (to assure positive numbers) and the natural logarithm is taken. The resulting distributions will be presented for each feature individually in the following subsections. Pattern recognition principles will then be applied to combine these features in a Bayesian classifier, assuming normal distribution of the component features.

3.1. Features for Stress Detection

Energy, loudness and amplitude have all been said to correlate with the percept of stress. Several experiments [9] were conducted to determine what the most useful measure of amplitude could be to obtain best separability between stressed and unstressed syllables, including *average* peak-to-peak amplitude over the extent of the *sonorant portion* of each syllable, the *average* peak-to-peak amplitude over the extent of the *entire* syllable, the *maximum* peak-to-peak amplitude over the extent of the *entire* syllable and the *integral* of the peak-to-peak amplitude over the extent of the *sonorant portion* of each syllable. For the syllables in our training database, the best separability was achieved by the integral of the peak-to-peak amplitude over the sonorant portion of the syllable, in good agreement with other studies [8]. Fig. 1 shows the PDFs for the logarithm of this integral for both the stressed and the unstressed syllables. The dotted curves show the approximation of the PDF by a Gaussian distribution (with mean and variance estimated from the training data).

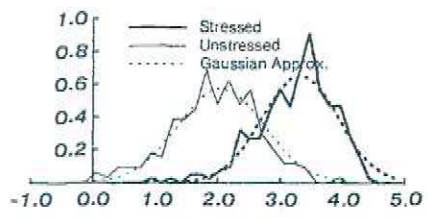

Figure 1. PDF of peak-to-peak amplitude integral over Sonorant Segments

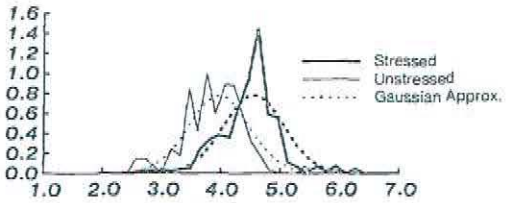

Figure 2. PDF of Syllable Duration

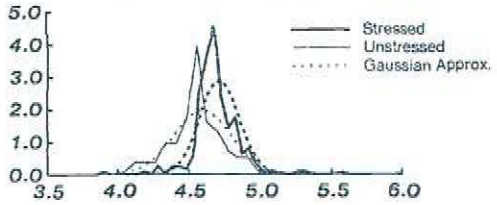

Figure 3. PDF of Pitch Maxima

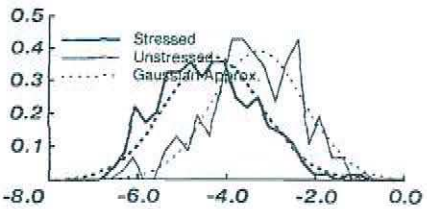

Figure 4. PDF of Average Spectral Change

Three separate measures of duration were taken: (1) the duration of the sonorant portion of the syllable in question, (2) the duration between syllable boundaries defined by the onsets of the sonorants of consecutive syllables[1], and (3) the duration between syllables defined by the onset of the consonant (cluster) preceding the syllabic nucleus in question. The best separability was obtained by the second measurement, i.e., the durations between syllables defined by the onset of their syllabic nuclei [9]. The PDFs corresponding to this measure are shown in Fig. 2.

To provide the pitch measurements, a feature based time domain pitch tracker [11] was used. A post-processing clean-up routine was necessary to remove irrelevant variability in the original pitch contour due to pitch tracker failures, segmental effects at consonant boundaries, or vocal fry. In this routine, all pitch values obtained in unvoiced regions were first set to zero. Next, pitch values with more than twice or less than half the average fundamental frequency are considered unreliable (pitch tracker errors or vocal fry) and are ignored. Spurious outliers are then removed using 5-point median smoothing. Finally, we wished to express the fall/rise movements at the syllabic level undisturbed by segmental effects (as, for example, encountered after the release of stop consonants), short pitch dips or peaks, or overall declination. Towards this goal, a best-fit linear approximation of the pitch contour in each syllabic segment was computed using a linear regression. Care was taken to exclude pitch pulses during the first 15 msec of a syllable to avoid perturbations due to segmental effects. In addition, a linear

[1] This measure approximates the duration between syllabic beats [10]

regression over the pitch values of an entire utterance was computed as a measure of declination [12]. All measures that were tried as features for stress detection were extracted from these short syllabic line segments and normalized by overall declination. Extensive experimentation suggests that this provides robust estimates of the local pitch variations in continuous speech [9]. Stress is typically accompanied by an increase in pitch as well as a rising and in some cases also a falling contour. Several measures might be suitable to capture the essential ingredients of stress. The specific measures of pitch that were tested include (1) the *slope* and the absolute value of the slope, (2) the *maximum value*, (3) the average value, and (4) the offset between pitch values from the previous syllable to the onset of the current syllable. The most promising measures were found to be the pitch maxima and (only slightly inferior) the pitch offset. Fig. 3 shows the PDF for pitch maxima.

Finally, average spectral change has been reported elsewhere as a useful measure for stress detection [5]. This is based on the observation that stressed syllables will be dominated by steady state vocalic portions rather than by short, rapidly changing, transitory vowels. As a measure, spectral change was averaged over the sonorant portion of a syllable nucleus. The resulting distributions are shown in Fig. 4.

3.2. Combination of Features - Optimal Stress Detection

We will now attempt to combine the features we have discussed so far in an optimal classifier. Based on the distributions obtained in the previous sections it is reasonable to assume normal (Gaussian) distributions of the features. The general approach, therefore, is to use a classical minimum probability-of-error decision rule [13], where we assume that the joint probability density function of the features (the measures discussed above) for a class i (in our case, i = 0,1 (stressed or unstressed)) is a multidimensional Gaussian distribution with known mean vector μ_i and covariance matrix Σ_i. Let x be an N-dimensional vector representing the measurements we have discussed above as its elements. Then the N-dimensional Gaussian density function is given by

$$p(x) = (2\pi)^{-N/2}|\Sigma_i|^{-1/2} \exp\left[-1/2(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)\right] \qquad (1)$$

Assuming this distribution, minimum-error-rate classification can be achieved using the discriminant function [13]

$$g_i(x) = -1/2(x-\mu_i)\Sigma_i^{-1}(x-\mu_i) - N/2 \log 2\pi - 1/2 \log|\Sigma_i| + \log P(\omega_i) \qquad (2)$$

The decision rule which minimizes the probability of error is that the sample described by the feature vector x should be assigned to the class i, which maximizes $g_i(x)$. In our simple case, i = 0,1 (stressed or unstressed), and our decision rule simply says that a syllable with feature vector x is called stressed if $g(x)_{stressed} > g(x)_{unstressed}$; otherwise it is called unstressed. In our case, Equation 2 can be simplified. First, we assume that stressed and unstressed syllables are equally likely to occur. This assumption holds true for our training data. The biasing *a priori* probabilities $P(\omega_i)$ will therefore be the same for both classes, i.e., 0.5. Second, the feature vector x describing a given syllable will have the same dimensionality N, regardless of class. Thus, since our decision depends only on the *relative* magnitude of the discriminant functions $g_i(x)$, we can rewrite Equation 2 as

$$g_i(x) = -(x-\mu_i)\Sigma_i^{-1}(x-\mu_i) - \log|\Sigma_i|$$

In this fashion, we achieve a combination of the features described in the previous sections into one classifier aimed at optimal decision between stressed and unstressed syllables. Probability of stressedness can be computed according to Equation 1. In the next section we will report a series of experiments that investigate various combinations of features and report the resulting recognition accuracies.

3.3. Performance Evaluation

Ptpint	Syldur	F0max	AveSpch	Missed	Extra	Ave.Err.
x				10.98%	18.03%	14.42%
	x			28.54%	12.79%	20.86%
		x		18.36%	46.75%	32.21%
			x	26.35%	19.08%	22.80%
x	x		x	09.98%	15.30%	12.58%
x	x	x	x	08.98%	15.72%	12.27%

Table 3-1: Stressed Syllable Classifiers Using Various Features - (test) database #4

Each of the measures described above was evaluated in the framework of a minimum error classifier. Detailed performance results of various combinations of features and detailed error analyses are described elsewhere [9]. We will here only summarize performance results using combinations of the most useful measures. Four databases were used for evaluation:

1. The first is the training database itself, which consists (as previously described) of 50 Harvard sentences, read by five different speakers (3 male, 2 female), each reading 10 different sentences. This database yields a total 482 syllables, 244 of which were determined by hand-labeling to be stressed and 238 to be unstressed.
2. The second database used is the Email task, which consists of 6 sets of 8 sentences each (a total of 48 sentences). Each sentence was read by a different speaker. Due to poor recording conditions 6 sentences are unusable, leaving 42 sentences for the evaluation. These 42 sentences resulted in 431 syllables, 202 of which were determined by hand-labels to be stressed and 229 to be unstressed.
3. The third database is given by two separate readings of the same 50 Harvard sentences we had used for training (with each speaker again reading 10 sentences). This database was generated and hand-labeled at the MIT Research Laboratory of Electronics. It presents data produced by different speakers and under different recording conditions than those used in training. This database yields 959 syllables, 488 of which were determined to be stressed and 471 to be unstressed.
4. The fourth database is similar to the third database described above. It was also generated at MIT and consists of 50 sentences read by 10 different speakers, 10 sentences each. The 50 sentences used here, however, are different from those in the first (training) and third databases. In addition, as in the third database, the speakers and the recording conditions differ from the database used for training. This database provided 978 syllables. This includes 501 stressed and 477 unstressed syllables.

Table 3-1 summarizes the results for the fourth database, a testing database, using different sentence material, speakers and recording

environment than those used in training. To arrive at these performance measures, a stress probability of 0.5 was arbitrarily set as the decision criterion between stressed and unstressed classes. Errors are tabulated in percent missed (all syllables that were labeled stressed but recognized as unstressed), percent extra (all syllables labeled unstressed, but recognized as stressed) and their average. The x-marks indicate the specific feature used in a given classifier. It can be seen that the best performance result was achieved by a classifier using all four features. Use of additional combinations of features such as the alternate measures mentioned above caused deterioration of results [9]. Performance results for the other databases (not shown here) were qualitatively similar. It is interesting to note that out of the four single-feature classifiers the peak-to-peak integral over the sonorant region of the syllable clearly performs best, and in fact independently achieves a respectable error rate (14.42%). This result is in good agreement with those of other studies [12, 9]. It indicates that peak-to-peak integral is perhaps the most powerful feature for discrimination between stressed and unstressed syllables. It might also have been predicted from the good separability between distributions found in Figure 1. All three additional features, however, do yield significant improvements over the one-feature classifiers.

Table 3-2 shows the individual error rates achieved by the best classifier for the various databases used. As could be expected, the error rate obtained by the training data is lowest. The classifiers perform worst on the Email task due to syllabification errors, and possibly due to emphatic stress (which tends to de-emphasize all but one prominent syllable). Perhaps not surprisingly, a positive correlation of .39 was found between the amplitude integral and duration feature in the training data. Also, duration and average spectral change showed a slight negative correlation of -0.22.

	DB-1	DB-2	DB-3	DB-4	Ave.
Missed	8.20%	14.85%	10.04%	8.98%	10.52%
Extra	8.40%	16.16%	17.41%	15.72%	14.42%
Ave.	8.30%	15.55%	13.66%	12.27%	12.44%

Table 3-2: Error Rates of the best classifier for each database

Extensive error analysis was performed for each error in the above experiments [9]. The two most prominent causes of error were syllabification errors and discrepancies between the actual realization of a syllable and the assigned lexical stress. The first situation occurs when the syllable boundary detection algorithm occasionally misses a boundary (e.g., "beauTY-OF"), or erroneously inserts a boundary (e.g., when a pop is detected as a voiced segment). A missed syllable boundary leads to long syllables recognized as stressed, while an extra boundary leads to detection of an additional (typically) unstressed syllable. The second major source of errors includes syllables with very short nuclei (e.g. "CHICKS", "KITTEN") leading to missed errors or syllables uttered longer than predicted lexically (e.g., the conjunction "AND" drawn out between clauses). Missed errors in short stressed syllables occurred more frequently in syllables with high vowels, since no phonetic information was taken into account for amplitude normalization. Many of the errors in this category, however, were indirectly due to a disagreement between the assumed model of lexical stress and the actual pronunciation. Stress is inherently a rather subjective percept, with human listeners also disagreeing about what is heard as stressed. In a separate experiment with two sets of seven human listeners [9] it was found that the correlation between stress votes of these two human groups (0.86) is about as good as between the human stress votes and this algorithm's stress probabilities.

In summary, this classifier appears to be robust, its features are relatively simple to compute, and it leads to consistently good recognition performance for all four databases. The present results exceed the performance achieved by other speaker independent continuous speech stress detection algorithms. In a forced decision task, error rates between 8.3% and 15.5% are obtained. Rather than making absolute decisions, the algorithm returns a probability of stressedness. The stress information obtained can be used in various ways in a continuous speech understanding system. It is a good predictor of the function/content word distinction (<15% error) with potential implications for lexical search and syntactic analysis. Other applications are currently being explored.

1. W.A. Lea, *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
2. D.H. Klatt and K.N. Stevens, "Sentence Recognition from Visual Examination of Spectrograms and Machine-Aided Lexical Searching," *Proceedings 1972 Conference on Speech Communication and Processing*, IEEE and AFCRL, 1972, pp. 315-318.
3. D.P. Huttenlocher and V.W. Zue, "A Model of Lexical Access Based on Partial Phonetic Information," *ICASSP '84 Proceedings*, IEEE, 1982, pp. 26.4.1-26.4.4.
4. A.Waibel, "Towards Very Large Vocabulary Word Recognition," Tech. report 144, Carnegie-Mellon University Computer Science Department, 1982.
5. A.M. Aull and V.W. Zue, "Lexical Stress Determination and its Application to Large Vocabulary Speech Recognition," *ICASSP '85 Proceedings*, IEEE, 1985, pp. 41.1.1-41.1.4.
6. I. Lehiste, *Suprasegmentals*, MIT-Press, Cambridge, MA, 1970.
7. M. Beckman, "Effects of Accent on Vowel Amplitude in Japanese," *The Journal of the Acoustical Society of America*, Vol. 75, 1984, pp. S41, (abstract only).
8. P. Lieberman, "Some Acoustic Correlates of Word Stress in American English," *The Journal of the Acoustical Society of America*, Vol. 32, No. 4, April 1960, pp. 451-454.
9. A. Waibel, "Automatic Detection of Lexical Stress", in preparation.
10. G.D. Allen, "The Location of Rhythmic Stress Beats in English: An Experimental Study I and II," *Language and Speech*, Vol. 15 and 16, 1972, pp. 72-100 and 179-195.
11. M.S. Phillips, "A Feature Based Time Domain Pitch Tracker," *The Journal of the Acoustical Society of America*, April 1985, , (abstract only).
12. P. Lieberman, W. Katz, A. Jongman, R. Zimmerman, M. Miller, "Measures of the Sentence Intonation of Read and Spontaneous Speech in American English," *The Journal of the Acoustical Society of America*, Vol. 77, No. 2, February 1985, pp. 649-657.
13. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, 1973.