Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing

Anders Eriksson, Gunilla C. Thunberg and Hartmut Traunmüller

Department of Linguistics
Stockholm University, Sweden
anders@ling.su.se, gunilla@ling.su.se, hartmut@ling.su.se

Abstract
In this experiment, subjects had to rate the “prominence” of each of the syllables of 20 versions of the same utterance produced by men, women and children at various levels of vocal effort. The ratings were correlated with measurements of the SPL of the fundamental, spectral emphasis, vowel duration, \( F_{\text{max}} \) and \( F_0 \) rise from the previous syllable. Together with ratings of the perceived vocal effort at which the utterances had been produced, these measurements were used to obtain the possible contributions of vocal effort, prosodic distinctness, and vowel duration to the perceived prominence. Together, these accounted for half of the variance. This was compared with the possible contribution of the linguistic structure of the utterance, which accounted for slightly more of the variance. The predictions of a model based on this analysis came closer to the mean than the average subject.

1. Introduction
In 1959, Lehiste and Peterson [1] suggested as an hypothesis that “the perception of linguistic stress is based upon judgments of the physiological effort involved in producing vowels”. Most subsequent analyses were, nevertheless, only concerned with easily measurable acoustic variables, such as SPL, \( F_0 \) and segment durations. Duration and intensity of vowels had already been shown to be correlated with stress in English bisyllabic words of the type in which stress placement is distinctive [2]. However, these acoustic variables provide sufficiently reliable cues for stress only in cases where they are not simultaneously used to signal other phonological distinctions. Higher pitch and larger pitch movements are also clearly associated with increased prominence of words and syllables [3, 4]. While this may be a fairly reliable correlate of prominence in Dutch, this is not the case in closely related languages such as in German and Danish, where it occurs that stressed syllables have a low \( F_0 \) throughout. Nevertheless, it may well be true for all languages in which stress has a communicative function that the potential pitch range is increased in all stressed syllables, albeit this is not evident in the signal when \( F_0 \) is low. However, in many languages, pitch (tone) and duration (quantity) of vowels are used for distinctions unrelated to stress, and the level of syllable nuclei varies with vowel quality and with non-linguistic factors.

The contribution of various features of the \( F_0 \)-contour of a sentence to the perceived prominence of its syllables has been investigated in detail, whereby attempts to scale \( F_0 \) in proportion with prominence [3, 4, 5] and to derive the prosodic baseline, against which the size of pitch excursions appears to be judged by listeners, have been in focus [4]. Streefkerk, Pols and ten Bosch [6] asked listeners to mark all stressed syllables in a large sample of sentences spoken by many different speakers. Subsequently they tested the goodness of several acoustical properties as predictors of prominence, whereby degree of prominence was equated with the number of subjects who had marked the word as stressed. They found clear correlations with both the median \( F_0 \) and \( F_0 \)-range (in st) as well as with syllable duration and with the calculated loudness of the vowel, expressed in relation to the average of the sentence. There was also a correlation with spectral slope, but this was weaker. On the face of it, this contrasts quite sharply with the result obtained by Campbell [7]. However, the data presented by Streefkerk et al. [6] are confounded by a probably very large degree of between-vowel variation in their measure of spectral slope, while Campbell [7] performed his analysis in a vowel-specific way.

There is substantial between-vowel variation in level (L), but less variation in the level of the first partial (\( L_0 \)). A convenient measure of spectral emphasis is then given by \( L - L_0 \). Unlike other measures of emphasis that have been used [6, 7, 8], this is not affected by variation in \( F_0 \) as such, but it is affected by between-vowel variation. Overall intensity (L) and this measure of emphasis were shown to be fairly reliable acoustic correlates of focal accents in Swedish [9].

If the prominence of syllables reflects the physiological effort involved in producing them, prominence will be correlated with the vocal effort at which the syllables were produced. In order to investigate this, we made use of speech material that had been used in an investigation of the acoustic effects of variation in vocal effort brought about by varying the distance between speaker and addressee [10], using the average rating of this distance obtained from listeners for each utterance [11] as a measure of vocal effort.

Most of the acoustic variables that have been shown to correlate with prominence (or stress) correlate also with vocal effort. However, increased pitch range is not a concomitant of increased vocal effort. Therefore prominence can not be a matter of vocal effort alone. It is also a matter of prosodic and articulatory distinctness. The latter is evidenced in the fact that stressed vowels show more extreme formant frequencies than the same vowels in unstressed position, but languages differ in the degree to which variation in stress affects the articulatory distinctness of vowels [12, 13]. The present experiment does, however, not address this question.

In order to investigate to what extent perceived syllable prominence can be understood as a function of variation in vocal effort between syllables, an experiment was designed in which subjects had to rate the prominence of syllables in a set of recorded sentences. These ratings were then correlated with acoustic variables known to be relevant for the description of vocal effort. It is to be expected, however, that the obtained
ratings also reflect aspects of prominence that are not due to vocal effort, but to prosodic distinctness and other factors.

In many investigations of prominence perception, subjects rated prominence just on a binary scale. However, listeners have been shown to be able to distinguish many more levels of prominence. In an experiment by Fant and Kruckenberg [14] subjects were instructed to indicate by pencil marks on vertical lines above the text the perceived stress magnitude of syllables in recorded sentences presented to them. The scale ranged from 0 to 30 where a value of 10 was to be considered typical for unstressed and 20 for stressed syllables. Before the listening test, however, subjects were told to rate “their own inner speech, when reading the text”. The ratings obtained in this way were closely similar to those obtained when listening to the reading of the text by a professional speaker. This is an indication that listeners may, to a considerable extent, depend on their own “top-down” interpretation in a rating task that involves real speech. This possibility will be further considered in the analysis of the results presented here.

2. Method

Eighteen adult speakers of standard Swedish (9 female, 9 male) served as subjects. All were employees or undergraduate students at the department.

The speech material was selected from recordings made for an investigation of the acoustic effects of variations in vocal effort [10]. It consisted of twenty utterances, recorded outdoors, in an acoustically free field in an area without disturbing noise. The utterances were of identical linguistic structure and content: Jag tog ett violett, åtta svarta och sex vita, ‘I took one purple, eight black and six white’, spoken at various degrees of vocal effort. The speakers were three men, three women, and four children (two boys and two girls), seven years of age. Each speaker was represented by two utterances produced at different vocal efforts.

The speech material was presented via headphones and the judgments were made on a computer screen, by shifting the positions of 13 sliders, one for each syllable, on a graphical display designed to look like a small sound mixer panel (see Fig. 1).

There was no response time limitation. The subjects could decide for themselves how many times to replay an utterance, and how much time to devote to adjusting the sliders for each utterance. A training session, using one utterance, preceded the test in order for the subjects to get acquainted with the use of the response tool.

Subjects were instructed to judge the “prominence” of each syllable within the utterance, one utterance at the time. To neutralize any possible between-stimulus contextual effects, presentation order was randomized, and different for all subjects. They were encouraged to use the whole range of possible positions for the sliders, placing one in top position for the most prominent syllable in the utterance (translated to 100%), as well as leaving one in the bottom position (0%) for the least prominent syllable. Despite the instructions, some subjects failed to make use of the whole scale. In these cases, the raw data were normalized linearly to agree with the provision.

The basic acoustic measurements were the following: fundamental frequency $F_0$, signal level $L$, fundamental level $L_0$, vowel duration. $L_0$ was defined as the level of the signal after low-pass filtering at 1.5 $F_0$ (-3 dB), with continuous adjustment of the cut-off frequency of a 4th order Butterworth filter. Emphasis was defined as $L-L_0$. The formant frequencies $F_1$ and $F_2$ were also measured, with moderate ambitions concerning accuracy, but with elimination of analysis frames in which the LPC-based automatic formant tracking procedure used produced obvious gross errors. In the subsequent analyses, pitch was expressed in semitones and the formant frequencies were also used in terms of their logarithms. Also vowel durations were considered in terms of their logarithms.

In a previous investigation, these same utterances had been presented to listeners who had to rate the distance between the speaker and the addressee [11]. In the present investigation, the mean values of those ratings were used as a measure of vocal effort [10]. Specifically, the 2-logarithms of the estimated distances in meters were used.

3. Results

The mean prominence ratings obtained from all subjects for each one of the syllables are plotted in Fig. 2 for each utterance. The three lines shown have been fitted to the mean data obtained from utterances whose communicational distance was estimated as less than 1.55 m, intermediate and more than 8.1 m (144, 90 and 126 utterance judgments, respectively). Syllables 1, 2, 3, 5, 6, 8, 11, and 12 may carry a main accent, syllables 7, 9, and 13 secondary accent. Syllables 3–9, and 11–13 are in words (the numerals and color terms) used contrastively within the sentence. The vowels in syllables 1, 2, and 12 are phonologically long. The sequence “io” in “violett” was also considered a single, long, diphthongal segment.

Figs. 3 to 7 show the acoustic measures taken. The between-syllable variation in emphasis was less pronounced in

![Figure 1](image1.png)

*Figure 1.* The response tool used by the subjects for rating the prominence of each syllable.

![Figure 2](image2.png)

*Figure 2.* Prominence ratings of the syllables. Mean values of all listeners' ratings.
utterances produced at a low vocal effort, but there was no such tendency in the prominence ratings.

While there was no obvious general variation as a function of overall vocal effort in any of the other acoustic variables, there was a tendency of reduced between-syllable variation in the first half of the utterances produced at a high degree of vocal effort. Towards the end of the same utterances, the variation was, instead, increased. This appears to be reflected also in the prominence ratings.

4. Data analysis and discussion

As a preparatory step, a linear regression analysis was performed, using the original L0, emphasis and F0mean as independent variables and the estimated communicational distance from [11] as the dependent variable (log. units). This resulted in a correlation coefficient of r = 0.991.

Using the regression equation obtained above, a new variable “apparent relative vocal effort” was calculated for each vowel on the basis of L0 (dB), emphasis (dB) and F0max (st). Here and in the following, all levels, segment durations, and frequency values were considered in relation to the mean of all vowel segments in the utterance. Since (L0 – L0) varies substantially between vowels produced at a given vocal effort, the calculated “apparent relative vocal effort” is substantially confounded by vowel quality. This is largely a not quite linear function of between-vowel variation in log(F1) and log(F2).

In a first linear regression analysis of the data, the “apparent relative vocal effort”, log(F1), log(F2), and their products with relative emphasis were used as independent variables, while the mean prominence rating for each syllable of each stimulus was used as the dependent variable. This resulted in a multiple r = 0.57. Without the correction for formant frequency effects, “apparent relative vocal effort” only explained 15% of the variance, but this value increased to 25% when the interaction between emphasis and vocal effort was taken into account. However, when F1 and F2 were used, an addition of the interaction factor produced only a negligible improvement. This means that the addition of the formant information accounted for this interaction as well.

In a second linear regression analysis, the following independent variables were used: (a) the pitch maximum of each vowel, in semitones above the average of all the vowels of the utterance; (b) the rise in pitch in semitones from the mean of the preceding syllable (For the initial syllable of the utterance and for syllables after pauses, variable (a) was taken as a sub-
stitute.), (c) the ordinal number of the syllable within the utterance; and (d, e, f) the products of the variables (a), (b) and (c) to account for interactions. The dependent variable was the mean prominence rating obtained for each syllable of each stimulus. This analysis was intended to capture the contribution of “prosodic distinctness” to perceived prominence. All variables (a) to (f) gave highly significant contributions. A rise in pitch has been suggested to be a strong stress cue for Swedish [15], but this has been questioned [16]. The present results suggest it to be a highly unreliable cue. The multiple r obtained was 0.51. The significance of the interactions (d) and (e) had been expected on the basis of the results reported in [4], who observed the contribution of pitch to prominence to vary with position in the sentence.

In a third linear regression analysis, the following variables were used: (a) the logarithm of the quotient between the duration of the vowel of a syllable and the mean duration of all vowels of the utterance; (b) a factor that was equal to one for syllables in pre-pausal position and zero elsewhere; (c) the product of (a) and (b) to capture possible interactions. The dependent variable was again the mean prominence rating for each syllable of each stimulus. All variables gave a highly significant contribution, with decreasing weight from (a) to (c). The multiple r obtained was 0.47.

The equations obtained in the preceding three analyses were used to calculate three summary variables: “vocal effort factor”, “pitch factor” and “duration factor”. These were used as independent variables in a further analysis, which resulted in a multiple r = 0.69 (48% explained variance). In this analysis, the weights of the independent variables were directly comparable. They were 0.70 for “vocal effort factor”, 0.54 for “pitch factor” and 0.49 for “duration factor”. These figures, which are roughly proportional to the variances explained, 33%, 26%, and 22%, could be taken as indicative of the relative importance of these signal based cues. Adding vowel quantity as an additional variable did not essentially affect the result. This may be due to the restricted speech material used.

An additional linear regression analysis concerned linguistic top-down factors. The independent variables, which had the values 1 and 0 for “yes” and “no”, respectively, were (a) syllable capable of carrying a main accent, (b) syllable capable of carrying a secondary accent and (c) syllable in word contrastively used within the sentence (numerals and color terms). The dependent variable was again the mean prominence rating. All the independent variables produced significant contributions. The multiple r obtained was 0.75, (57% explained variance). Although this is more than the 48% explained by signal-based cues, this needs not mean that prominence perception is mainly a top-down process.

5. Conclusions

The results show that subjects can use vocal effort, the distinctness of F0-movements, and vowel duration as cues for rating syllable prominence. However, we can not tell which cues they actually used. A strategy based mainly on top-down processing could have produced a similar result. However, the success of prominence predictions based on the variables “vocal effort factor”, “pitch factor” and “duration factor”, was quite high, although these accounted for less than half of the variance in the data. The average error of the prominence values predicted by a model based on these factors was 18.2 units, which is markedly lower than the standard deviation of the subjects’ ratings (24.5 units). Thus, the model can be said to be substantially better than the average human subject.

6. Acknowledgments

This research is supported by a grant from HSFR, the Swedish Research Council for the Humanities and Social Sciences.

7. References