Perception of Syllable Prominence by Listeners with and without Competence in the Tested Language

Anders Eriksson[1], *Esther Grabe*[2] *& Hartmut Traunmüller*[1]

[1]Department of Linguistics, Stockholm University, Sweden
[2]Phonetics Laboratory, University of Oxford, UK
anders@ling.su.se

Abstract

In an experiment reported previously, subjects rated perceived syllable prominence in a Swedish utterance produced by ten speakers at various levels of vocal effort. The analysis showed that about half of the variance could be accounted for by acoustic factors. Slightly more than half could be accounted for by linguistic factors. Here, we report two additional experiments. In the first, we attempted to eliminate the linguistic factors by repeating the Swedish listening experiment with English listeners who had no knowledge of Swedish. In the second, we investigated the prominence pattern Swedish subjects expect by presenting the utterance only in written form. The results from these subjects and from the Swedish listeners were very similar but for two of the syllables where the prominence pattern did not coincide with the expectations of the readers. Swedish and English listeners perceived the prominence of the syllables to be almost identical in most cases, but where there was a conflict between expected and produced prominence, the Swedish listeners appeared to be influenced by their expectations. There was also a difference in the weights the Swedish and English listeners attached to different acoustic cues in the listening experiments.

1. Introduction

The production, perception and comprehension of prominence distinctions in speech varies across languages (e.g. [3], [10], [14], [15]). In English, stressed syllables provide segmentation points for words, but in Japanese, stressed syllables are not exploited for this purpose [1], [13]. And some languages are said to be produced with a stress-timed rhythm whereas others are syllable-timed or mora-timed [2], [9]. Consequently, if prominence judgments or rhythmic classifications are made by native speakers from different languages, we can be faced with apparently conflicting results. Miller [12] asked English and French subjects to classify rhythmically eight languages. The results provided support for the traditional classifications of languages only in one case: Arabic. Miller's study raises the following questions:

(1) What prominence patterns do native and non-native listeners hear when they judge the rhythm of a language?
(2) How do their prominence perceptions correlate with the acoustic structure of the stimuli?

In 1959, Lehiste and Peterson [11] 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 [6]. Higher pitch and larger pitch movements are also clearly associated with increased prominence of words and syllables [7], [16]. However, these acoustic variables provide sufficiently reliable cues for stress only in cases where they are not simultaneously used to signal other phonological distinctions.

In many investigations of prominence perception, subjects have rated prominence 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 [5] 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. 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.

In order to investigate to what extent perceived syllable prominence can be understood as a function of variation in vocal effort between syllables, Eriksson *et al*. [4] designed an experiment 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.

2. Method

2.1. Speech material

The speech material was selected from recordings made for an investigation of the acoustic effects of variations in vocal effort [17]. 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 in response to the question *Hur många kort tog du av varje färg*? 'How many cards of each colour did you take?' 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.

2.2. Response collection

The speech material was presented via headphones and judgments were made on a computer screen, by shifting the positions of a number of sliders 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. A training session, using one utterance, preceded the test in order for the subjects to get acquainted with 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 effects, presentation order was randomised, 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.

2.3. Acoustic measurements

The basic acoustic measurements were the following: fundamental frequency F_0, signal level L, fundamental level L_0 and 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 [18]. In the present investigation, the mean values of those ratings were used as a measure of vocal effort [4]. Specifically, the 2-logarithms of the estimated distances in meters were used. A linear regression analysis was performed, using the original L_0, emphasis and F_{0mean} as independent variables and the estimated communicational distance [18] as the dependent variable (log. units). This resulted in a correlation coefficient of r=0.991. Using the regression equation obtained in this way, the "apparent relative vocal effort" was calculated for each vowel on the basis of L_0 (dB), emphasis (dB) and F_{0max} (st).

2.4. Experiment 1

In Experiment 1, eighteen adult speakers of standard Swedish (9 female, 9 male) served as subjects. All were employees or undergraduate students at the Department of Linguistics at Stockholm University. The subjects judged the prominence of each syllable of the utterance.

2.5. Experiment 2

This experiment was carried out at the University of Oxford in the UK. The speech material was the same as that used in Exp.1. Ten adult speakers (5 female, 5 male) of Southern

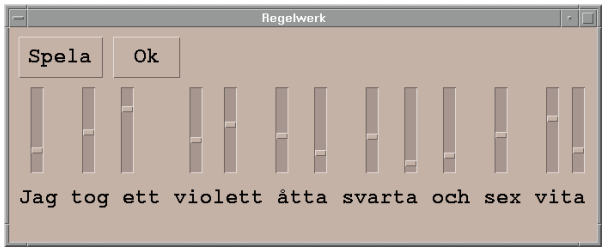

Figure 1. The magnitude estimation tool used by the subjects for rating the prominence of each syllable.

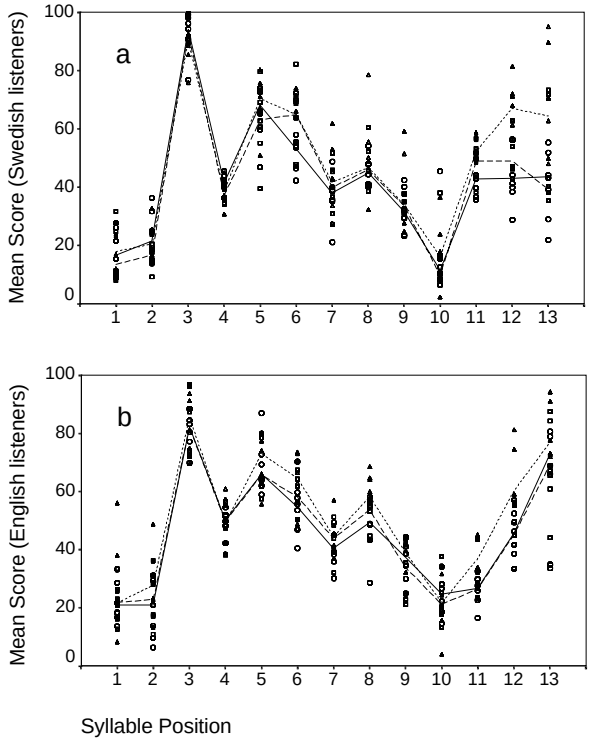

Figure 2 a–b. Prominence ratings of the syllables. Mean values of all listeners' ratings. The lines represent mean values, syllable by syllable, for the three levels of vocal effort used to produce the stimuli. Dotted line: high, solid line: intermediate, broken line: low.

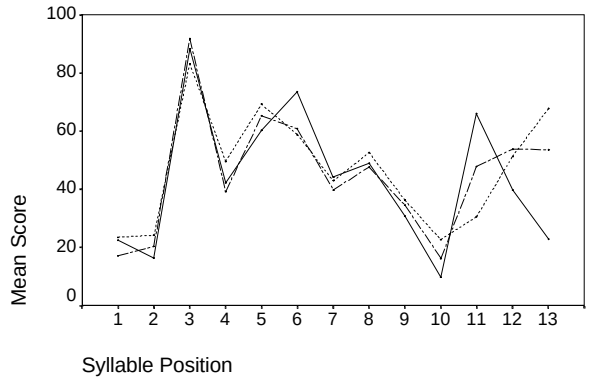

Figure 3. Mean prominence ratings for the three experiments. Broken line: Swedish listeners, dotted line: English listeners, solid line: Swedish non-listeners.

British English with no knowledge of Swedish took part in the study. All subjects were undergraduates at the University of Oxford.

2.6. Experiment 3

Eighteen adults (12 female, 6 male) with the same background as the those taking part in Exp. 1 served as subjects. To this group of subjects, the utterance was only presented in written form and they were told that it was a response to the question used in recording the stimuli. They were asked to imagine how they would produce the utterance and then indicate the prominence level of each syllable using the magnitude estimation tool.

3. Results and discussion

The mean prominence ratings obtained from all subjects for each one of the syllables are plotted in Fig. 2a (Swedish listeners, Exp. 1) and 2b (English listeners, Exp. 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). Fig. 3 shows the mean ratings of the prominence expected for each syllable by speakers of Swedish (Exp. 3) together with the mean auditory judgments by the English and Swedish listeners, without effort distinction.

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

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 $(L - L_0)$ 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 $\log(F_1)$ and $\log(F_2)$.

In a first linear regression analysis of the data obtained in Exp. 1 (Swedish listeners) and 2 (English listeners), the "apparent relative vocal effort", $\log(F_1)$, $\log(F_2)$, 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 (Swedish) and 0.56 (English).

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 substitute.), (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 [10], but this has been questioned [8]. The present results suggest it to be a highly unreliable cue. The multiple r obtained was 0.51 (Swedish) and 0.57 (English). The significance of the interactions (d) and (e) had been expected on the basis of the results reported in [7], 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 0.48 (Swedish) and 0.58 (English).

The equations obtained in the preceding three analyses of the two experiments 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 (Swedish) and 0.75 (English). (48% and 56% explained variance, resp.). In this analysis, the weights of the independent variables were directly comparable within as well as between the two experiments. They were 0.70 (Swedish) and 0.60 (English) for "vocal effort factor", 0.54 (Swedish) and 0.58 (English) for "pitch factor" and 0.49 (Swedish) and 0.56 (English) for "duration factor". These figures, which are roughly proportional to the variances explained, 33%, 26%, 22%, (Swedish) and 32%, 33%, 33% (English) could be taken as indicative of the relative importance of these signal based cues.

The correlation coefficient obtained between the ratings of the Swedish readers (Exp.3) and listeners (Exp. 1) was 0.77. That obtained between the ratings of the Swedish readers (Exp. 3) and the English listeners (Exp. 2) was also significant, 0.59. Since the English result could only be based on acoustic properties, this tells us that the linguistic properties must to a large extent have been encoded in the acoustic signal.

The results show that subjects are able to use vocal effort, the distinctness of F_0-movements, and vowel duration as cues for rating syllable prominence. The success of prominence predictions based on the variables "vocal effort factor", "pitch factor" and "duration factor", was quite high, although these accounted for just about half of the variance - somewhat more in the English than in the Swedish data. The average error of the prominence values predicted by a model based on these factors was 16.4 units (Swedish) and 13.9 (English), which is markedly lower than the standard deviation of the subjects' ratings, 24.5 units (Swedish) and 24.3 (English). Thus, compared with random selection of a human subject as a representative of the behaviour of his group, the models based on acoustic analysis produce a substantially better description. Despite these promising results, we can not tell which weight the Swedish listeners actually attached to the different acoustic cues. This is due to the fact that Exp. 3, in which the subjects had to rely entirely on top-down processing, produced a similar result.

The difference between the results of the Swedish and the English listeners can, in part, be understood as due to interference, among the Swedes, of their a priori expectations. These are reflected in the results of the silent experiment (Exp. 3). Such interference can only show itself when there is

a discrepancy between the a priori expected and the actually realized prominence of a syllable. This explains why the acoustic cues explain more of the variance in the results of the English subjects as compared with the Swedish: in the absence of any a priori expectations (English subjects) there is no source of such interference.

In Fig. 3, it can be seen that the discrepancy between a priori expectations (Swedish non-listeners, Exp. 3) and acoustic realizations (English listeners, Exp. 2) was most pronounced in syllables 11 (less prominent than expected) and 13 (more prominent than expected). Here, the interference shows itself clearly in the fact that the results of the Swedish listeners (Exp. 1) deviate from those of the English listeners in the direction indicated by the results of the Swedish non-listeners (Exp.3).

We can also see that the Swedish listeners attached relatively more weight to vocal effort, while the English attached about equal weight to effort, pitch and duration. This has probably to do with the fact that in Swedish, pitch and duration are used for additional phonological distinctions, while the English listeners can be assumed to have based their judgements on English conditions.

4. Conclusions

The present experiments have shown that listeners are able to judge the relative prominence of syllables in an utterance by making use of the variation in a number of acoustic factors. Based on these factors, a model was suggested which was able to explain more than half of the observed variance in the prominence ratings by the subjects. For most syllables, Swedish and English listeners judged prominence to be the same. A cross-linguistic difference could be observed, however, where there was a conflict between the actual realization of a syllable and the linguistically based expectations by listeners competent in the tested language.

However, the speech material used in the experiment had been recorded for a different purpose and was not ideally suited for the present type of study. In order to obtain clearer results, it is necessary to vary the speech material in a more systematic fashion by introducing more variation that is not predictable from the linguistic structure, for example the location of focus within an utterance. In future work, we will compare prominence ratings from a wider range of languages including French (French speakers are well known to have difficulties with stress distinctions in English [15]) and Finnish where quantity distinctions occur in both stressed and unstressed syllables in contrast to Swedish where quantity distinctions only occur in stressed syllables.

5. Acknowledgments

This research is supported by a grant from HSFR, the Swedish Research Council for the Humanities and Social Sciences (H. Traunmüller, A. Eriksson) and the UK Economic and Social Research Council (E. Grabe).

6. References

[1] Cutler, A.; Norris, D., 1988. The role of strong syllables in segmentation for lexical access. *J. Exp. Psychol. Human*., 14, 113–121.

[2] Dauer, R. M., 1983. Stress-timing and syllable-timing reanalyzed. *J. Phonetics*, 11, 51–62.

[3] Dupoux, E.; Peperkamp, S. in press. Fossil markers of language development: phonological 'deafnesses' in adult speech processing. In J. Durand & B. Laks (eds.) *Phonetics, Phonology, and Cognition.* Oxford: Oxford University Press, 168–190.

[4] Eriksson, A.; Thunberg, G. C.; Traunmüller, H., 2001. Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. *Proc. EUROSPEECH '01*, Vol. 1, 399–402.

[5] Fant, G.; Kruckenberg, A., 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*, 2, 1–83, Stockholm: KTH.

[6] Fry, D. B., 1955. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.*, 27, 765–768.

[7] Gussenhoven, C.; Repp, B. H.; Rietveld, A.; Rump, H. H.; Terken, J., 1997. The perceptual prominence of fundamental frequency peaks. *J. Acoust. Soc. Am.*, 102, 3009–3022.

[8] Heldner, M.; Strangert, E., 1997. To what extent is perceived focus determined by F_0-cues? *Proc. EUROSPEECH '97*, Vol. 2, 875–878.

[9] Hoeqvist, Jr., C., 1983. Syllable duration in stress-, syllable- and mora-timed languages. *Phonetica*, 40, 203-237.

[10] House, D.; Hermes, D.; Beaugendre, F., 1998. Perception of tonal rises and falls for accentuation and phrasing in Swedish. *Proc. ICSLP '98*, Vol. 6, 2799–2802.

[11] Lehiste, I.; Peterson, G. E., 1959. Vowel amplitude and phonemic stress in American English. *J. Acoust. Soc. Am.*, 31, 428–435.

[12] Miller, M., 1984. On the perception of rhythm. *J. Phonetics*, 12, 75–83.

[13] Otake, T.; Hatano, G.; Cutler, A.; Mehler, J., 1993. Mora or syllable? Speech segmentation in Japanese. *J. Mem. Lang.*, 32, 258–278.

[14] Peperkamp, S.; Dupoux, E.; Sebastián-Gallés, N. 1999. Perception of stress by French, Spanish, and bilingual subjects. *Proc. EUROSPEECH '99*, Vol. 6, 2683–2686.

[15] Peperkamp, S.; Dupoux, E. in press. A typological study of stress 'deafness'. In C. Gussenhoven; N. Warner (eds.) *Laboratory Phonology 7*. Berlin : Mouton de Gruyter.

[16] Rietveld, A. C. M.; Gussenhoven, C., 1985. On the relation between pitch excursion size and prominence. *J. Phonetics*, 13, 299–308.

[17] Traunmüller, H.; Eriksson, A., 2000. Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am.*, 107, 3438–3451.

[18] Rundlöf, J., 1996. *Perceptuella ledtrådar vid auditiv bedömning av avståndet mellan talare och lyssnare*, Stockholm: Department of Linguistics, Stockholm University.