ISCA Archive
http://www.isca-speech.org/archive

4th International Symposium on
Tonal Aspects of Languages (TAL-2014)
Nijmegen, The Netherlands
May 13-16 2014

Acoustic correlates and perceptual cues of word and sentence stress: Mainly English and Dutch

Vincent J. van Heuven

Leiden University Centre for Linguistic, Leiden, The Netherlands
Dept. Applied Linguistics, University of Pannonia, Veszprém, Hungary
v.j.j.p.van.heuven@hum.leidenuniv.nl

Abstract

This is an abridged version of a tutorial on the (older) experimental literature on the acoustic correlates of, and their contribution to, human stress perception, mainly on the basis of research on English and Dutch. The emphasis is on establishing the relative importance of correlates amd cues. The conclusion is that a reliable acoustic correlate (e.g. peak intensity) is not necessarily a strong perceptual cue. Conversely, the strongest perceptual cue (i.e. pitch change) is acoustically an unreliable correlate. The full tutorial will be accessible at https://openaccess.leidenuniv.nl/search?query=heuven.

Index Terms: word stress, sentence stress, acoustic correlate, perceptual cue

1. Introduction

The purpose of this paper is to present and discuss the way word and sentence stress are phonetically marked. It has been known since the 1950s that stress (whether at the word or sentence level) is never marked by a single acoustical property (for a survey see [1]). To make the stressed syllable stand out from its neighbours, it is produced with greater physiological effort on the part of the speaker than its unstressed counterpart (e.g. [2]). The greater effort will be exerted at any stage in the speech production process, i.e., by the subglottal mechanism (more air is pushed out of the lungs), by the glottal system (contraction of laryngeal muscles, generating a change in pitch) and by the supraglottal organs (e.g. larger and faster displacement of lips, tongue and jaw, yielding more clearly articulated vowels and consonants). The greater effort is seen, first of all, in closer approximation of articulatory target configurations for segments in stressed syllables. More extreme articulatory movements require more time than small displacements of the vocal organs. The result of this is that segments in stressed syllables have longer durations – all else being equal – than unstressed segments.

We will not deal any further with the physiological basis of stress. We will concentrate on the acoustic consequences of increased versus decreased effort and ask (i) what acoustic correlates can be found for the difference between a stressed syllable and its unstressed counterpart, and (ii) what the relative importance is of each acoustic correlate in the marking of stress. At the same time we will consider the question what acoustic properties are used by human listeners and to what extent these are used to decide whether or not a syllable is stressed. We will make a strict terminological distinction here between acoustic correlates of stress (which can be used, for instance, to identify a stressed syllable by some computer algorithm) and the perceptual cues used by the human listener. We will see that some acoustic correlates, notably the (peak) intensity of a syllable, allow good separation of stressed from unstressed tokens but are hardly used by the human listener.

2. Acoustic correlates

It is generally not a good idea to just compare acoustic properties of successive syllables in a word. If the segmental make-up of the syllables is different, the correlates of stress are obscured by the intrinsic and co-intrinsic properties of the segments. For instance, open vowels have inherently greater intensity [3] and longer duration [4] than close vowels so that an unstressed open vowel may, in fact, seem more stressed than a closed stressed vowel, as may happen in the English noun ***IM**pact*. Several tricks have been suggested to eliminate, or correct for, such inherent segmental properties. One way out would be to use so-called reiterant speech [5, 6, 7], where the speaker replaces the syllables in a target word by repetitions of the same segmental structure, e.g. of /ma/ or /lis/ (e.g. the target utterance *please say **IM**port again* would be produced as *please say **MA**ma again*, or *please say **LIS**lis again*). The claim is that the speaker dubs all (and only) the prosodically relevant variations onto the reiterant version of the original utterance so that no segmental normalisation is needed. A potential problem with these techniques is that stressed and unstressed syllables are compared syntagmatically, i.e. in different linear positions in a larger structure, such as an initial stressed and a final unstressed syllable – so that it remains unclear whether we measure correlates of stress or of sequential position. The safest precaution, therefore, would be to compare stressed and unstressed versions of the same syllables in a paradigmatic way, e.g., by comparing the stressed and unstressed realisations of the first and second syllables in a minimal stress pair such as *the IMport* versus *to imPORT*. This solution only works if the language has at least one minimal stress pair – it cannot be used in languages with fixed stress.

It has also been found expedient to measure the correlates of stress separately for stress at the word level and at the sentence level. This is generally achieved by (paradigmatically) comparing tokens of stressed and unstressed syllables in a minimal stress pair which was produced in the same position in a surface-syntactically identical sentence, with and without focus on the target. Focus on the target word (indicated in 1a-d in square brackets) is often manipulated by having the speaker answer different questions that highlight one constituent or the other (sentence stress in capitals):

(1a) Q: Did you read 'the import' or 'the sale' again?
A: I read ['the IMport'] again
(1b) Q: Did you read 'to import' or 'to sell' again?
A: I read ['to imPORT'] again
(1c) Q: Did you read 'the import' again or write it down?
A: I [READ] 'the **IM**port' again
(1d) Q: Did you read 'to import' again or write it down?
A: I [READ] 'to im**PORT**' again

We will now briefly review what has been reported in the literature on the acoustical marking of word and sentence stress. I will mainly draw on publications on Dutch and Eng-

lish. We will begin by discussing properties that are found equally in word and sentence stress and finish by zooming in on those properties that differentiate word from sentence stress (and are found, therefore, only when a syllable occurs in a word with sentence stress).

Temporal organisation. Since the work by Fry [8] it has been clear that stressed syllables – all else being equal – are longer than their unstressed counterparts. Fry measured the duration of the first and second vowels (V_1 and V_2) in five English minimal stress pairs (noun-verb pairs *contract*, *digest*, *object*, *permit*, and *subject*) spoken once by twelve American speakers in sentence-final position in a fixed carrier *Where is the accent in…*, which elicits sentence stress on the target words. With the duration of V_1 and V_2 as predictors, a Linear Discriminant Analysis (LDA [9]), a classification algorithm often used for this purpose, yields 83% correct classification of stress pattern (computed by me (VH) from data in Fry's appendix). After z-normalising V_1 and V_2 duration within stress pairs, percent correct classification of stress pattern rises to 93. Next, we may apply intrinsic normalisation by computing the relative duration of the first vowel (V_1%) as a percentage of the summed durations of V_1 and V_2). We then find just one single case in which V_1% was the same for the noun and the verb reading of the pair; in all other 59 cases V_1% was larger for the noun (initial stress) than for the verb (final stress) reading (98% correct classification). The conclusion is that vowel duration is a very good correlate of stress. Fry ([8]: 765), however, remarks that consonant duration ratios were 'not materially affected by the shift of stress'. This conclusion deserves further scrutiny. I turn to data on Dutch to examine effects of stress on subsyllabic units, i.e. vowels, onset and coda consonants separately.

An early study that examined the effect of stress on the durations of subsyllabic units in Dutch can be found in [10: appendices 11-12). Target items were non-words /papapap/ and /papapap/, with short/lax /ɑ/ and long/tense /a/. Items were spoken with stress on the first, second and third syllable in turn, in carrier sentences such that they were either 'accented' (with sentence stress) or 'unaccented' (word stress only). A large number of tokens were produced by two male Dutch speakers. The results are summarized in Figure 1.

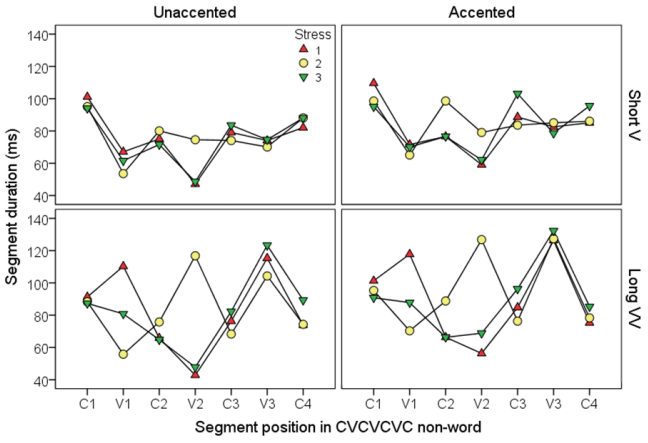

Figure 1. *Segment duration (ms) in the sequence /pVpVpVp/ as a function of stress position (initial, medial, final) in unaccented vs. accented Dutch non-words with short (lax) and long (tense) vowels (data from [10], appendices 11-12).*

The relative effects of stress on the temporal make-up of the non-words are very similar for accented and unaccented items – although durations are consistently longer overall under sentence stress. Hardly any effects of stress can be seen in the final syllable. There are very large differences in the durations of V_1 and V_2 depending on the stress position. When the item is spoken with initial stress, V_1 is very long and V_2 short (ratio $V_1/V_2 > 1$). With medial stress, this pattern reverses completely, with a very short V_1 and a very long V_2 (ratio < 1), while items with final stress have intermediate vowel durations for V_1 and V_2 (ratio ≈ 1). The crucial observation, however, is that the effect of stress position on the durations of the consonant segments, though small in absolute terms, appears to be quite consistent as well: it is nearly always the case that a C, whether onset or coda, is somewhat longer on average in the stressed version of the syllable than in the unstressed version (i.e. in a paradigmatic comparison).

Intensity. Intensities of speech sounds are unstable as they vary considerably (intensity drops in the order of 5 dB), e.g., when the speaker inadvertently turns his head. Intensity differences of similar magnitude have commonly been reported as correlates of stress. These differences are small but prove reliable correlates (i.e. with little variability) of sentence stress but are even smaller and less reliable when word stress is signaled (cf. [11, 12] for English; [13, 14, 15, 16] for Dutch). In all these (and other) studies peak intensity was measured, which is usually reached shortly after the vowel onset.

In a paradigmatic comparison, i.e. comparing the stressed and unstressed reading of the same vowel in the same position in minimal stress pairs, the stressed version in Fry [8] had more decibels than the unstressed counterpart in 52 out of 60 V_1 pairs and in 55 V_2 pairs. Moreover, it is nearly always the case that the intensity difference between V_1 and V_2 was more positive in the noun reading (with stress on V_1) than in the correponding verb reading (with stress on V_2). Out of 60 comparisons 58 behaved as predicted, in one case the relationship was reversed and in one more the noun and the verb reading had the same intensity difference between V_1 and V_2. This makes (peak) intensity, and especially intensity difference between stressed and unstressed syllables a very reliable acoustic correlate of stress in English. It should be pointed out in this context that [8] is often misquoted. It is not the case that his data show that intensity is a poor *acoustic* correlate of stress or that it is a poorer correlate than duration.

Spectral balance. Stress in Western Germanic languages has often been equated with the expenditure of vocal effort, which is correlated with perceived loudness. The most obvious acoustic correlate of physiological effort and perceived loudness, it was held, is vocal intensity. Increased pulmonary effort causes a larger volume-velocity of airflow through the glottis. The result is not just the generation of larger glottal pulses but also, and more importantly, of a more strongly asymmetrical glottal pulse (Figure 2).

The closing phase of the glottal period is shortened, yielding a smaller opening quotient (OQ, i.e. the proportion of the time the glottis is open relative to the period duration T), and the trailing edge of the glottal pulse is steeper. The greater steepness of the glottal closure and its abrupt ending (smaller Closure Quotient, CQ), cause the generation of relatively strong higher harmonics, with a flatter spectral tilt (Figure 3).

The effects of stress on spectral tilt at the sentence (left-hand column) and word level (right-hand column) can be seen in Figure 4 for a paradigmatic comparison of selected syllables in the Dutch minimal stress pair *CAnon ~ kaNON* /ˈkanɔn ~ kaˈnɔn/ 'round song ~ cannon' and reiterant mimicry by five male and five female speakers. Generally, no effects of stress

can be observed in the base band (< .5 KHz). Effects are strong in the higher frequency bands, causing flatter spectral tilt, especially under sentence stress, and more clearly so in the initial syllable than in the final syllabe.

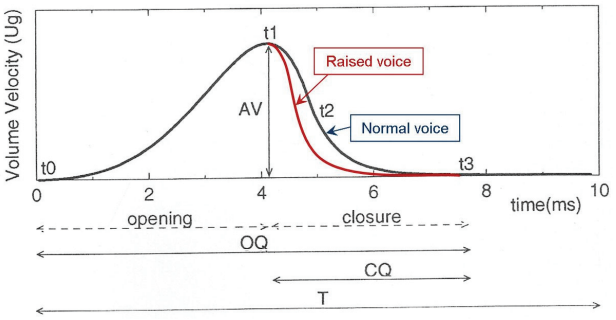

Figure 2. *Effect of normal versus raised voice on volume-velocity of airflow through glottis. t1: maximum flow during glottal cycle, t2: fastest decrease of glottal flow, t3: complete glottal closure (no flow). Graph based on [16]*

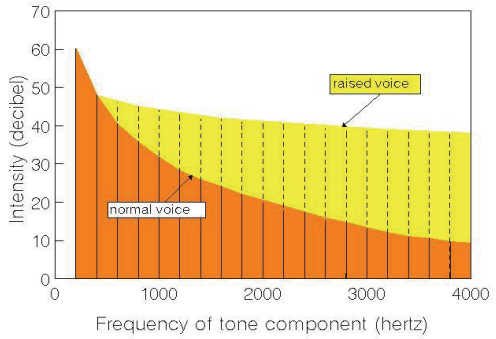

Figure 3. *Effect of decreased Open Quotient (OQ) and Closure Quotient (CQ) due to raised voice on the spectral envelop (difference is exaggerated).*

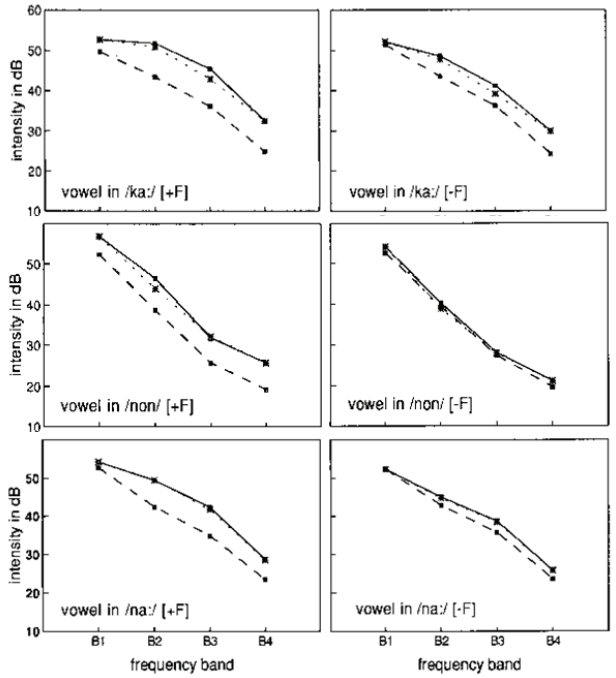

Figure 4. *Effects of sentence (left column) and word (right column) stress on spectral tilt. Intensity (in dB) is plotted for four frequency bands (B1: < .5 KHz, B2: .5-1 KHz, B3: 1-2 KHz, B4: 2-4 KHz). Further see text.*

Spectral expansion. Stressed vowels have often been described as 'clear' (spectrally expanded), reflecting greater articulatory effort and precision. These vowels lack the spectral reduction that is typical of unstressed vowels. Figure 5 (based on [17]) illustrates the effects of word and sentence stress on the expansion/reduction of long (tense) Dutch /e:, o:, a:/ read by 15 male speakers. The position of the schwa (averaged over 300 tokens across consonant environments and speakers) serves as the centre of gravity of the vowel space.

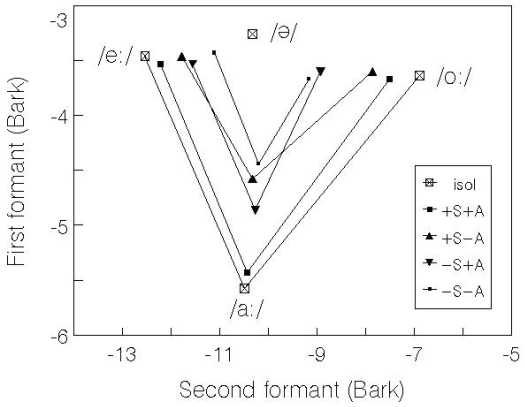

Figure 5. *F_1 and F_2 (Bark) of three Dutch tense peripheral vowels produced by 15 male speakers in five stress conditions (see text, after [17]).*

Spectral expansion is largest for vowels pronounced in isolation ('isol'). Some reduction is visible when these vowels occur in the stressed syllable of accented words ('+S+A' = sentence stress). Considerable reduction is seen for stressed vowels in unaccented words ('+S−A' = word stress) or for unstressed vowels in accented words ('−S+A'). Severe spectral reduction is found in unstressed vowels of unaccented words ('−S−A'): here the spectral distance to /ə/ is minimal. Similar results were obtained for reiterant American English non-words by [15] (for details see [16]: 116-117).

Automatic classification of stress by spectral expansion of Dutch vowels was done by [15] in the minimal stress pair /ˈkanɔn ~ kaˈnɔn/ (see above) and their reiterant versions (/nana/) produced in a short carrier with and without word and sentence stress (four combinations). Predictors in the LDA were the F_1 and F_2 of V_1 and V_2. Percentages of correct stress identification were 84 and 77 for words with and without sentence stress, respectively, and 68 and 71 for the reiterant non-words. These identification scores are better than chance (= 50%) but are poorer than what was observed for most other stress correlates (see below).

Acoustic correlates of sentence stress. As long as there is no sentence stress on a word, the speaker makes no effort to change the vocal pitch. To be true, there may well be a small rise-fall contour on any vowel (with or without word stress) but this is due to an involuntary response of the glottal mechanism to the greater transglottal pressure that comes about when the oral tract opens during the articulation of the vowel sound; during the articulation of consonants the oral tract is fully or partially closed so that intraoral impedance yield a transglottal pressure drop causing the vocal folds to vibrate more slowly. It has been estimated that the involuntary effect of mouth opening on the rate of vocal fold vibration does not normally exceed a threshold of 4 semitones (a frequency rise and subsequent fall of less than 25 percent). Only when a word is produced with sentence stress does the speaker issue a voluntary command to the glottal muscles that brings about a change in pitch greater

than 4 semitones. Listeners intuitively know that smaller changes in vocal pitch require no planned action on the part of the speaker and therefore ignore these as a stress cue.

For a pitch change to impart sentence stress on a syllable the change has to be strictly local, i.e. has to take place within a time window that does not exceed the duration of a syllable. Gradual pitch movements (rises or falls than span a longer sequence of syllables) can never be prominence lending [18]. Yet, not any large and fast change in vocal pitch is associated with sentence stress. Fast pitch changes may also be used to mark prosodic boundaries. The difference between prominence-lending and boundary-marking pitch changes is in their timing relative to the segmental structure of the syllable. In Dutch, for instance, an equally large and fast pitch rise located in the first half of a syllable imparts prominence (sentence stress) but it marks the syllable as domain-final (intonation domain boundary or question marker) rather than stressed when executed in the final portion of the syllable (end of rise aligned to end of voicing).

Data collected by [19] (see [16]: 106-116 for a more extended report) illustrate the point. Three male and three female speakers of American English each recorded two tokens of four minimal stress pairs (the noun-verb pairs *export*, *uplift*, *digest* and *compact*) as well as their reiterant versions with syllables /bi/, /bɛ/ and /bɑ/, medially in fixed carrier sentences such that targets received either sentence stress or not. The f_0 change under sentence stress was two to three times larger (in semitones) than in items with word stress only. Most of the f_0 movements associated with word stress only were below 4 semitones. When the token was produced with sentence stress it was nearly always the case that the f_0 peak fell within the stressed syllable affording perfect identification of stress pattern in the four lexical pairs and near perfect stress identification in the reiterant versions (98% correct). However, when tokens were produced with word stress only (with phrase-final sentence stress), the locations of the f_0 peak were distributed more evenly over the two syllables and were aligned with the stress in only 65% of the cases (chance = 50%).

Relative strength of stress correlates. Using the LDA automatic classification algorithm as an estimator of effect size, the number of (above chance) classification errors serves as a good approximation of the relative strength of an acoustic correlate of stress. We had the LDA classify initial and final stressed members of reiterant minimal stress pairs produced with and without sentence stress by six native speakers of American English [20], separately for word stress (targets outside focus) and sentence stress (targets in focus). Predictors were in both conditions: (i) the location of the F_0 peak (in first or second syllable), (ii) relative duration of the first syllable, (iii) difference in peak intensity between the syllables, (iv) the difference in Euclidean distance of the vowel from the centre of the formant space, and (v) the difference between the syllables on an index based on five glottal parameters (not discussed here).

F_0, duration and intensity afforded very good classification of stress pattern for sentence stress (above 95% correct), vowel quality yielded only 80% correct classification. The estimated glottal source parameters afforded between 69 and 79% correct classification (the latter for spectral tilt between fundamental and F_2), with an exception of amplitude of the fundamental, which yielded 97% correct and was in fact slightly better as a predictor than just overall peak intensity). Much poorer classification was obtained for word stress (in words out of focus). Location of the F_0 peak, intensity, OQ and amplitude of fundamental were all between 60 and 65% correct (chance = 50%). B_1 and the two tilt measures were at 75% correct. The best classification was given by duration and vowel quality (both at 80%).

A provisional conclusion from this comparison of parameter strengths would be that the difference between initial and final stress is more clearly marked in English when it is a matter of sentence stress than when we are dealing with just word stress. The effect sizes of the parameters differ substantially between sentence stress and word stress. The location of the F_0 peak, peak intensity and amplitude of the fundamental are strong correlates in the sentence stress condition but not for word stress. Duration is a reliable correlate in both conditions, and so is spectral quality – be it less reliable than duration. Spectral tilt measures are only moderately successful correlates.

3. Perceptual cues

We will now review the perceptual cue value of the stress correlates discussed above. These studies compare the cue value of pairs of acoustic correlates in relatively small sets of stimuli. For instance, Fry published a series of three experiments comparing the strength of vowel duration (as a baseline condition) with that of three other parameters, viz. peak intensity [8], f_0 [21] and vowel quality [22]. This series of experiments should yield a rank order of perceptual importance for the four correlates.

Duration vs. intensity. Figure 6a (left-hand panel) shows the main results of the perception study by Fry [8]). In the experiments the durations of V_1 and V_2 in each of five minimal stress pairs (*object*, *subject*, *digest*, *compact*, *import*, see above) were varied in five steps between (and including) values found (averaged over ten speakers) in natural tokens with initial and with final stress. These five duration steps were combined with five intensity differences (by amplifying V_1 and at the same time attenuating V_2) such that the V_1–V_2 difference varied between +10 and –10 dB. Listeners indicated whether they perceived a noun (initial stress) or a verb (final stress). Unfortunately, Fry did not present the results for the individual stimulus types. Instead, Figure 6 (after Fry's Figure 3) presents percent perceived initial stress for duration steps (averaged over words and intensity steps) and for intensity steps (averaged over words and duration ratios).

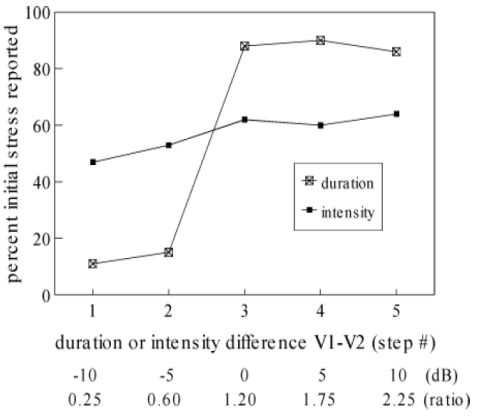

Figure 6. *Initial stress perceived (%) as a function of duration of V_1/V_2 and of intensity difference between V_1 and V_2 in English minimal stress pairs. (after [8])*

Figure 7 shows the results of a similar experiment for a single Dutch minimal stress pair, the reiterant non-word *nana* [21]. The results are practically the same as in English. However,

there are more and smaller stimulus steps, which makes the cross-over appear somewhat more gradual than Fry's. Also, the targets were presented in a sentence frame *wil je [target] ZEGgen* 'will you [target] SAY' with the sentence stress on the final verb; these variations were suggestive of word stress only – the range of intensity differences in the Dutch stimuli was much smaller (but reflected actual speech production) than in Fry's materials with sentence stress on the targets.

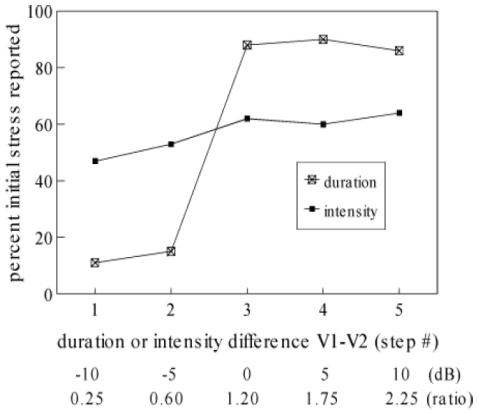

Figure 7. *As Figure 6a but for Dutch. (after [15])*

Figure 8 is a quasi-3D plot of percent initial stress perceived as a function of the difference in vowel duration (X-axis) and of the difference in intensity (Y-axis). The boundary in the figure separates the white area with a majority of initial-stress decisions from the dark area with a majority of final stress responses. In panel 8A the boundary runs at an angle that is much steeper than 45°, which indicates that the duration parameter outweighs the intensity parameter as a stress cue. It also shows that intensity variations are largely inconsequential: they cannot swing the majority decision from initial to final stress for six out of seven duration steps; only when V_1 = 170 ms and V_2 = 245 ms does intensity yield a (shallow) cross-over from 43 to 60% initial-stress responses.

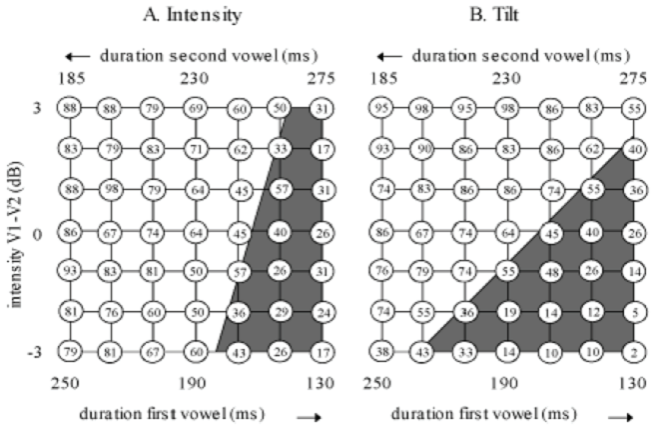

Figure 8. *Initial stress perceived (%) as a function of temporal structure (duration of V_1 and V_2, X-axis) and of intensity difference (Y-axis). A: uniform intensity variation (changing gain factor). B: intensity variation at frequencies > .5KHz only (yielding differences in spectral tilt).*

Duration vs. selective intensity (spectral slope). We also included a set of stimuli in which the same intensity differences were generated on V_1 and V_2 but in such a way that no differences were made at frequencies below 500 Hz and all the changes were concentrated at frequencies above 500 Hz, thereby creating a change in spectral slope [21]. Panel 8B shows that (selective) intensity differences (affecting spectral tilt) are as strong a stress cue as are the duration differences: the boundary now runs at a 45° angle. In this experiment, the stimuli had been presented over headphones with artificial reverberation added. The reverb (realistic of room acoustics) obscures temporal details. When the same materials were presented over headphones without reverb, the effects of selective intensity were smaller than those of duration but still larger than those of uniform intensity differences.

Contribution of consonant vs. vowel duration. Now that we have seen that duration generally outweighs other cues for word stress, let us examine the effects of the duration of subsyllabic units such as the onset consonant, the vocalic nucleus and the coda consonant. In reiterant stimuli, with short/lax vowels (/pafpaf, tastas/) and with long/ tense vowels (/pafpaf, tastas/), we varied the durations of onset, nucleus and coda separately in steps of 50, 75, 100, 125 and 150 percent of the original duration [22]. The stimuli were synthesized from diphones that had been excerpted from stressed syllables produced in nonsense words with sentence stress, so that all original segments were equally suggestive of (strong) stress. Figure 9 presents the results.

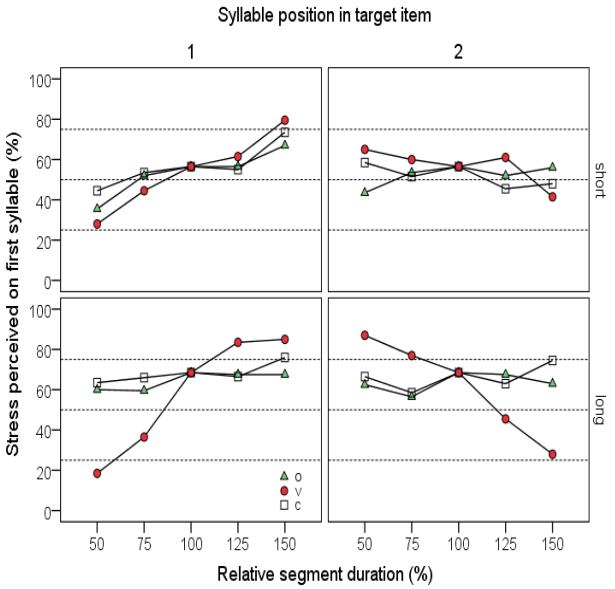

Figure 9. *Percent stress perceived on first syllable as a function of relative duration of onset, vocalic nucleus and coda in either first (left panels) or second (right panels) syllables with short/ lax (upper panels) or long/tense (lower panels) vowel.*

Figure 9 shows that, overall, effects of changing the duration of the vocalic nucleus are large but changes in consonant durations, whether in the onset or in the coda, have little or no effect on stress perception. A complete cross-over from stress perceived on the first syllable (S_1) to stress perceived on the second syllable (S_2) is found for vowel duration change, except when the vowel is short (lax) and in the final syllable of the target non-word (top-right panel). Moreover, the effect of changing the (vowel) duration is weaker overall when the changes are implemented S_2 than in S_1. Changing the duration of a consonant only affects stress perception if the change takes place in an S_1 with a short (lax) vowel (top-left panel) but even then the effect is still somewhat smaller for consonants than for the vowel. In this condition, it does not matter whether the consonant is in the onset or in the coda. So, it seems safe to

conclude that the older literature was right in assuming that vowel duration by itself, rather than syllable duration or rhyme duration, is the relevant duration cue.

Duration vs. vowel quality. The only study on the effect of vowel quality on stress perception in English was done by Fry [23]. Fry manipulated the formants of vowels in four stress pairs (*contrast*, *digest*, *object*, *subject*). Keeping pitch and intensity differences constant, the duration ratio and formant structure of V_1 and V_2 were varied in three steps each, creating a 3 × 3 = 9 item stimulus space for each noun-verb pair, i.e. 45 stimuli in all. Formants F_1 and F_2 in V_1 were manipulated for three words pairs (*contrast*, *digest*, *object*) while keeping V_2 constant; formants in V_2 were varied in *object* and *subject* while keeping V_1 constant. The formant manipulations were such that either F_1 or F_2 or both moved one step towards the centre of the vowel space (suggesting vowel reduction). Figure 10A plots duration and formant changes such that more initial stress should be perceived going from left to right. The results indicate that stress is less likely to be perceived on the syllable with reduced vowel quality; the tendency is somewhat stronger when the vowel quality is reduced in the F_2 dimension (backness and rounding) than in the F_1 dimension (height) and is strongest when both quality dimensions are affected simultaneously. Reducing vowel quality, however, does not yield a convincing cross-over: percent initial stress changes from 45 to 60. The effect of duration is clearly stronger.

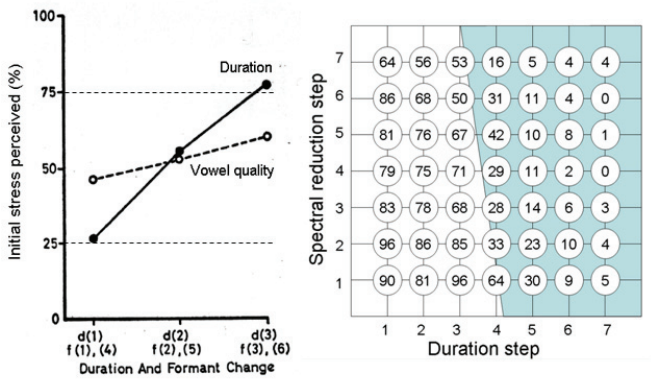

Figure 10. *A (left): Percent initial stress perceived in English as a function of V_1/V_2 duration ratio and of vowel reduction in F_1 (left), F_2 (middle) or both (right) in either V_1 (steps f1..f3) or V_2 (steps f4..f6; after [23]). B (right): Percent initial stress perceived in Dutch as a function of V_1/V_2 duration ratio and spectral reduction in V_1. (from [24])*

Fry [23] did not vary vowel quality in terms of an acoustic continuum. A more direct comparison of vowel duration and quality was made for Dutch [24]. We varied the V_1/V_2 ratio and the quality of V_1 in the Dutch stress pair *CAnon ~ kaNON* (see above) in seven steps along each continuum. Targets were presented in postfocal position (no f_0 movement on the target) in a carrier *ik heb GISteren een CAnon (kaNON) gehoord* /ɪk hɛp ['ɣɪstərən]$_{+F}$ ən 'kanɔn (ka'nɔn) ɣə'hɔrt/ 'I have yesterday a canon (cannon) heard', i.e. 'I heard a canon (cannon) yesterday'. The results are shown in Figure 10A, in quasi-3D format. Convincing cross-overs are obtained for the duration steps. Just one, very incomplete, change from perceived initial stress to final stress is obtained by changing vowel quality from clear to fully reduced to schwa; this change is obtained only when the duration cue is ambiguous (step 4). Fry's conclusion is confirmed here: vowel reduction is a much weaker stress cue than vowel duration.

Duration versus fundamental frequency. Let us, finally, examine the perceptual effects of varying the size and segmental alignment of f_0 changes as a cue for stress. As I pointed out earlier, in natural human speech the f_0 change has to exceed a certain threshold (say > 4 semitones) in order to function as a stress cue, and if it does it typically imparts sentence stress on the word that carries the f_0 change. Since sentence stress outranks word stress, this makes the f_0 change the strongest stress cue of all. Fry [25] was among the first to study the effect of f_0 change on stress perception, comparing its strength with that of varying the duration ratio of V_1 and V_2 in the English noun-verb pair *subject*. The duration ratio was varied as in [8]. In one experiment, Fry synthesized the syllable *sub-* on a flat 97 Hz followed by stepwise f_0 rise to *-ject* of 5, 10, 15, 20, 30, 40, 60 and 90 Hz. This set of eight rises was supplemented with a similar set of eight falls, with the level higher f_0 on *sub-* and the low 97 Hz pitch on *-ject*. The total set of 5 (V_1/V_2 ratios) × 8 (step sizes) × 2 (directions) = 80 stimuli. The results bear out that the frequency step-up generated perceived stress on the second syllable (between 61 and 75% for the various f_0 changes but averaged over duration ratios) whilst a step down yielded stress on the first syllable (between 48 and 80%), i.e. the higher-pitched syllable is heard as stressed. The absolute size of the step, however, did not matter: a 5-Hz change was as influential as a 90-Hz change. On average, however, the effect of changing f_0 turned out to be smaller than that of varying the duration ratio.

In a second experiment, Fry [25] combined the five vowel duration ratios with 16 different f_0 contours. Two f_0 contours always yielded initial stress, even if the duration ratio strongly suggested final stress. Three contours always yielded a majority of final-stress judgments. In the remaining eleven contours, however, there was always at least one duration ratio that could swing the stress from initial to final, thereby counteracting the effect of f_0. A possible interpretation of the results is that an f_0 change involving a properly aligned high target cannot be counteracted by any duration ratio. But even for the most extreme duration ratios there was always an f_0 pattern that could swing the judgments from initial to final stress. By this reasoning f_0 chance outranks duration as a stress cue.

Van Katwijk [13: 76-88] varied f_0 movements in a Dutch reiterant nonsense item /sœsœsœs/ in a rather realistic fashion. F_0 changes were implemented relative to a fixed declination of 5 st/s. Keeping all other parameters constant, f_0 rises and falls of 3 st during 100 ms were generated at eleven different time points. Table 1 specifies the alignment for the onset of the f_0 movement with repect to the duration of a segment. Here 'V_1 00' means that the f_0 movement begins at 0% of the duration of the first vowel, i.e. at the vowel onset. Van Katwijk [13] also generated three stimuli with rise-fall contours, and two (one rise, one fall) with 6-st excursion sizes (during 200 ms). The results show that the location of the f_0 movement greatly influences the perception of stress. A simple rise or rise+fall at the beginning of a syllable suffices to attract a clear majority of stress responses to that syllable (indicated by yellow shading in Table 1). Simple falls tend to attract fewer stress judgments than rises do, especially when they are associated with the medial or final syllable. For a simple f_0 fall to impart stress on a syllable it has to be aligned rather late in the syllable or even in the beginning of the next syllabe. The complex rise-fall does not attract more stress judgments than a simple rise; long 6-st rises and falls do not attract more stress judgments than 3-st exemplars. Van Katwijk [13] also generated stimuli with differences in vowel duration and intensity but never in combination with f_0, or with each other, so that no direct comparison of cue strengths is possible.

Table 1. *Number of (sentence) stresses perceived by 45 Dutch listeners (free choice) on S_1, S_2 and S_3 in the nonsense word* /sœsœsœs/. *Further see text. (after [13]: 81-83)*

Rise 3 st				Fall 3 st				Rise-fall 3 st			
align	S_1	S_2	S_3	align	S_1	S_2	S_3	align	S_1	S_2	S_3
V_1 00	**31**	6	9	C_1 50	22	5	6	V_1 00	**41**	2	6
V_1 25	**33**	10	4	V_1 00	**36**	4	1	C_2 75	9	**42**	3
V_1 50	13	**35**	6	V_1 50	**36**	4	8	C_3 50	5	4	**38**
C_2 00	8	**39**	10	C_2 25	**37**	8	6				
C_2 50	6	**44**	6	C_2 75	**35**	6	7	Rise 6st			
V_2 00	4	**37**	6	V_2 00	22	**18**	4	C_1 00	7	**44**	21
V_2 50	11	18	**32**	V_2 50	19	**17**	**18**				
C_3 00	16	4	**43**	C_3 25	20	9	**19**	Fall 6st			
C_3 50	14	2	**45**	C_3 75	17	4	**32**	C_2 50	**35**	20	24
V_3 00	16	3	**39**	V_3 00	12	3	**28**				
V_3 50	22	4	6	V_3 50	17	2	4				

Perceptual rank order of cues. The most important perceptual cue for stress (in English and Dutch) is a change in fundamental frequency (if properly aligned with the segmental structure). The second-most influential cue is temporal organisation, specifically the duration ratio between the stressed and the unstressed version of the vowels (rather than of the consonants). Intensity would seem to rank third, but only if it is implemented such that the gain or loss of intensity is concentrated in frequency bands above 500 Hz, thereby affecting the slope of the spectrum (the flatter the spectrum, the greater the perceived loudness). Overall intensity and vowel quality are the weakest cues (unclear which one would be weaker).

Fry (for English) as well as Van Katwijk (for Dutch) insist that f_0 change is a stronger stress cue than duration. This claim is rather unsubstantiated, however, either because the experiment does not allow the conclusion to be drawn, or because the crucial data were not presented. Although Fry [25] provides at least circumstantial evidence, it is not the case that an f_0 change cannot be overridden by temporal cues in his materials.

4. Conclusions

The most important conclusion is that the strength of acoustic correlates of stress and the perceptual cue value of these correlates are not rank-ordered in a one-to-one fashion. This has two reasons. First, the location of an f_0 change is a strong correlate of stress in speech production only if the f_0 change exceeds a threshold of 3 to 4 semitones and if it is appropriately aligned with the segmental structure. When words do not receive sentence stress, the f_0 change is no longer a reliable correlate. For f_0 change to be a perceptual cue, no such threshold is required: even a small change (from 97 to 104 Hz) is enough to evoke final-stress perception, while a fall of the same size yields initial stress. Therefore, f_0 change may be perceptually the strongest cue but it is acoustically unreliable. Second, listeners do not rely on uniform intensity differences. This makes intensity one of the weakest perceptual cues, even though it is acoustically quite reliable. The relative unresponsiveness of the human hearing mechanism to differences in intensity has, in fact, been known for over a century ([26, 27, 28]). Differences in vowel duration are both perceptually strong and acoustically highly reliable, both for word stress and sentence stress.

5. References

[1] Lehiste, I., Suprasegmentals, MIT Press, 1970.

[2] Ladefoged, P., "Stress and respiratory activity", in Three areas of experimental phonetics, Oxford University Press, 1-49, 1967.

[3] Lehiste, I. and Peterson, G. E., "Vowel amplitude and phoneme stress in American English", J. Acoust. Soc. Am., 31:428-435, 1959.

[4] Peterson, G. E. and Lehiste, I., "Duration of syllable nuclei in English, J. Acoust. Soc. Am., 32:693-703, 1960.

[5] Larkey, L. S., "reiterant speech: An acoustic and perceptual validation", J. Acoust. Soc. Am., 73:1337-1345, 1982.

[6] Liberman, M. Y. and Streeter, L. A., "The use of nonsense-syllable mimicry in the study of prosodic phenomena", J. Acoust. Soc. Am., 63:231-233, 1978.

[7] Nakatani, L. H. and Shaffer, J. A., "Hearing 'words' without words: prosodic cues for word perception", J. Acoust. Soc. Am., 63:234-245, 1978.

[8] Fry, D. B., "Duration and Intensity as physical correlates of linguistic stress", J. Acoust. Soc. Am., 27:765-768, 1955.

[9] Klecka, W. R., Discriminant analysis, Sage, 1980.

[10] Nooteboom, S. G., Production and perception of vowel duration, a study of durational properties of vowels in Dutch, doct. diss., Utrecht University, 1972.

[11] Lea, W. A., "Acoustic correlates of stress and juncture", in L. Hyman [Ed] Studies in stress and accent, SCOPIL, 4:83-119, 1977.

[12] Beckman, M. E., Stress and non-stress accent, Foris, 1986.

[13] Katwijk, A. van, Accentuation in Dutch; An experimental linguistic study, Van Gorcum, 1974.

[14] Rietveld, A. C. M., Syllaben, klemtoon en de automatische detectie van beklemtoonde lettergrepen in het Nederlands [Syllables, stress and the automatic detection of stressed syllables in Dutch], doct. diss., Catholic University of Nijmegen, 1984.

[15] Sluijter, A. M. C. and Heuven, V. J. van, "Spectral balance as an acoustic correlate of linguistic stress", J. Acoust. Soc. Am., 100:2471-2485, 1996.

[16] Sluijter, A. M. C., Phonetic correlates of stress and accent, HIL Dissertations, 15, Holland Academic Graphics, 1995.

[17] Bergem, D. van, "Acoustic vowel reduction as a function of sentence accent, word stress, and word class on the quality of vowels", *Speech Comm*., 12:1-23, 1993.

[18] Hart, J. 't, Collier, R. and Cohen, A., A perceptual study of intonation. Cambridge University Press, 1990.

[19] Sluijter, A. M. C., Shattuck-Hufnagel, S., Stevens, K. N. and Heuven, V. J. van, "Supralaryngeal resonance and glottal pulse shape as correlates of prosodic stress and accent in American English", Proc. 13th Int. Cong. Phon. Sc., Stockholm, 2:630-633, 1995.

[20] Sluijter, A. M. C. and Heuven, V. J, van,. "Acoustic correlates of linguistic stress and accent in Dutch and American English", Proc. ICSLP 96. Philadelphia: Applied Science and Engineering Laboratories, Alfred I. duPont Institute, 630-633, 1996.

[21] Sluijter, A. M. C., Heuven, V. J. van and Pacilly, J. J. A., "Spectral balance as a cue in the perception of linguistic stress", J. Acoust. Soc. Am., 101:503-513, 1997.

[22] Heuven, V. J. van, "Stress and segment duration in Dutch", in Kager, R., Grijzenhout, J. and Sebregts, K. [Eds] Where the principles fail. A festschrift for Wim Zonneveld on the occasion of his 64th birthday, Utrecht Institute of Linguistics OTS, 217-228, 2014.

[23] Fry, D. B., "The dependence of stress judgments on vowel formant structure", in Zwirner, E. and Bethge, W. [Eds] Proc. 6th Int. Cong. Phon. Sc., Karger, 306-311, 1965.

[24] Heuven V.J. van and Jonge, M. de, "Spectral and temporal reduction as stress cues in Dutch", Phonetica, 68:120-132, 2011.

[25] Fry, D. B., "Experiments in the perception of stress", Lang. Speech, 1:126-152, 1958.

[26] F. L. Saran, Deutsche Verslehre, C.H. Beck, 1907.

[27] Mol, H., Uhlenbeck, E. M., "The linguistic relevance of intensity in stress", Lingua, 5:205-213, 1956.

[28] Bolinger, D. L., "A theory of pitch accent in English", Word, 14:109-149, 1958.