Teaching rules for English rhythm/timing


1. The teaching of English rhythm to foreign learners (and speech-synthesizers?) is (i) crucial, but also (ii) neglected and under-developed.

This is quite obvious but not generally realised. The first point hardly needs comment. It is obvious when one listens to some Dutch/French/Indian/etc. accents of English, which can be difficult or even impossible to understand because of their un-English rhythm.

As to the second point, the only approach available to English teachers, to my knowledge, is the so-called ‘stress-timing’ approach. In my talk I tried to show that this does not work by saying The North Wind and the Sun (in section 7, below) with ‘isochronous stresses’. It makes one sound rather like a square-dancing master intoning ‘take your partner and walk along, hold her hand and step to the left, now turn around and …’, etc. Any such attempt at a regular, ‘isochronic’ rhythm would indeed be quite ridiculous in ordinary speech. So perhaps we should try a different, ‘allochronic’ approach.

2. Physically and phonetically, (English) rhythm is a matter of the timing of syllables only, not of pitch or loudness. “If you weren’t already convinced of this you may now agree that there is nothing wrong with my English rhythm, even though I am now speaking on a monotone giving every syllable the same loudness. It just makes me sound a little bit more boring than usual, that’s all.”

The previous paragraph was said without any pitch variations, i.e. all of it on the single note d = 147 Hertz, thereby making and proving the point concerned simultaneously.

By saying a text both ‘normally’ and on a monotone, it transpires that one cannot change its rhythm by changing the pitch or loudness, only by changing syllable durations. That is not to say that pitch and loudness play no part, but their rhythmic contribution is at the most suggestive or supporting, not crucial/defining. Music teachers, making a clear distinction between rhythm and pitch, have always known this, but linguists and phoneticians tend to mix up rhythm and pitch, notably when talking about ‘stress’.
3. Mentally and phonologically, rhythm is a more complicated matter of organising one’s syllables on (largely) semantic criteria into groups and groups-within-groups of **Strong(er)** syllables with or without 1 or (at the most) 2 **weak(er)** syllables between them.

This is largely done on communicative/semantic criteria, i.e. by organising one’s words into rhythm-groups or bytes, one’s bytes into tone-groups or pieces, and one’s pieces into breath-groups or locutions.

For example: your 'house' is on 'fire' - /jəˈhaus/ tə̃ vən 'faɪə/ (said either with or without pitch-variation) is one idea or piece of information consisting of 5 words. But saying it, I am not thinking of every single word. And pointing or gesturing with my hands, I am not making five physical gestures, but only two. Correspondingly, the whole piece consists of only two mental gestures or ‘thoughts’, not five. First I think of /jəˈhaus/, with *house* as a Specified, Strongly Stressed concept (an S-word), and *your* as an unstressed, unemphatic ‘automatic reflex’ (a u-word) not requiring any thinking or mental effort at all.

Then I think of tə̃ vən 'faɪə', with the words *is* and *on* again as automatic reflexes or u-words, and *fire* as a mental effort or gesture, an S-word.

So words combine into rhythmic bits or bites (spelt ‘bytes’ by really with-it scientists), consisting of one Strongly Stressed or S-word, with or without one or more unstressed or u-words. Such rhythm-groups of one S-word with or without automatic reflex words are also known as rhythm-waves (Pike), ‘groupes rythmiques’ (Passy), measures and bars. We’ll just call them bytes.

So /jəˈhaus/ tə̃ vən 'faɪə/ consists of 5 words, i.e. concepts, combining into 2 bytes, i.e. thoughts. And the two bytes combine into one single piece, i.e. a ‘constellation’ of thoughts = idea, like a constellation of stars in the sky.

If I add another piece to the piece /jəˈhaus/ tə̃ vən 'faɪə/ in the same breath, the two pieces together make up a locution, i.e. a constellation of ideas = proposition. E.g. your ‘house’ is on ‘fire’ if you ‘want’ to ‘know’

---

4. **General phonetic theory/terms needed:** 1 phonetic/psychological *second* = 12 *moras*; 1 *mora* = 2 to 3 *points*. One minute = 60 to 70 seconds.

In order to talk about rhythm, any phonetics teacher (like any music teacher) needs a theory+notation. Musicians divide larger units, equal, isochronous, ‘regular’ bars into whole, half, quarter, eighth, etc. notes or values.

Since in speech we fail to find such regular *isochronic* units (measures, bars, feet, rhythm-groups, bytes…), the timing of any such *allochronic* unit can
only be stated as the sum-total of its constituent syllables (= the musician’s ‘notes’).

In order to get round this problem I’ll divide the psychological or ‘phonetic’
second into 12 moras, i.e. approximately the smallest time-values people can
comfortably handle. The initial syllable durations of pe-pə, peɪ-pə, pæ-pə -
pepper, paper, pauper, are approximately 2, 3, 4 moras, i.e. differing by such a
minimal amount. For any further refinement, if needed, the mora may be said
to consist of 2 to 3 points. To allow for differences in people’s psychological
seconds it may be defined (as appropriate) as between a sixtieth and a
seventieth part of a minute. According to some, it corresponds to the maternal
heartbeat, imprinted on the mind during one’s sojourn in the womb.

The resulting numerical notation (shown in section 7) also seems more
convenient than traditional musical notation. It is not more ‘artificial’ or
counter-intuitive, as long as it is kept in mind that such numbers only indicate
fractions (albeit of seconds or heartbeats rather than of bars or measures).
Clearly, they do not, and should not, indicate absolute time-values like
deciseconds or centiseconds. Nevertheless, they do indicate approximately
absolute, reasonably accurate syllable-durations, unlike musical notes or
descriptive labels such as long, short, half-long and ‘isochronous’.

5. Musical notation equivalent to numerical notation.

So English syllable durations can be shown by numerical values: 8 moras
for the longest, down to 1 mora for the shortest possible syllable, according to
the timing rules given in the next section.

If one should still prefer a musical to a numerical notation, the following is
to show that our numerical notation is easily converted into an equivalent
musical notation. This also means that teaching rhythm ‘by numerical
notation’ is no different in principle from traditional (elementary) rhythm
teaching by music teachers.

1 mora = ¼ quaver = eighth-note 5 moras = ¾
2 moras = ¼ crotchet = quarter-note 6 moras = ½
3 moras = ½ minim = half-note 7 moras = ¾
4 moras = ½ semibreve = whole note 8 moras = ¾
6. The following four ‘Syllable Timing’ rules are prescriptive rules for teaching foreign learners of English. They are also meant to be descriptive, but of course only of one style in one idiolect.

ST1. Allocate 2/3/4 moras resp. to syllables containing short/medium/long V.

ST2. Add 1 mora for C (sequence) following V in the same syllable.

ST3. Lengthen syllables before byte\| piece\| loc\# boundary by 1/2/3 moras resp.

ST4. Shorten u-syllables by 1 mora.

ST1. To apply this rule, one has to know which vowels are short, medium and long in English. Roughly: the short vowels are those of pit, pet, pat, put, also o, and also those of ti fo tso if unstressed/written singly (e.g. in hu\#u2\#). The long vowels are those of dcct, h33t/q if stressed/written double) and meri f3ari [that] f3ari. All other vowels are medium. [Corresponding spelling-forms: pit, pet, pat, putt, pot, put; tea for two, who is she, daughter hurting father, Mary fearing his fury.]

ST2. To add a mora for a syllable-final consonant (sequence) one must of course know where to draw the syllable-boundary. This is briefly explained in section 7, note 2, below.

ST3 requires an understanding of the phonological units ‘byte, piece, locution’ with their symbols |, \#, \, their corresponding forms ‘rhythm-group, tone-group, breath-group’ and their corresponding meanings ‘thought, idea, proposition’. Cf. the discussion in section 3 above. ST3 is crucial as it brings meaning into the specification of English syllable-durations.

ST4. u(nstressed) syllables are those having z(ero) or w(eak) stress, i.e. not having S(trong) or M(edium) stress. Cf. section 7, note 1.

This rhythm ‘calculus’ covers nearly all the ground and is reasonably easy to teach and apply. Two further ‘refinement’ rules will complete the picture, providing the (very subtle) finishing touches. The rules are obviously not descriptive of all rhythms of all English speakers, but rather of one style in one idiolect (like any description, say, of English vowel articulations).


Note: 2 degrees of ‘stressed’: S(trong), M(edium); 2 degrees of ‘unstressed’ w(eak), z(ero). The ‘normal’ 2 degrees primary=S and secondary=M are indicated in the normal manner by raised and lowered stress-marks. I also distinguish two degrees of ‘unstressed’: minimal=z-stress and non-minimal=w-stress. Again following normal practice, z-stress is left unmarked, but w-
stresses (required by the Rhythmic Alternation Principle implied in section 3) are marked with a small dot, e.g. the syllable \textit{don} in the second locution \textit{stronger| don di mo#} and the syllables \textit{wiz} and \textit{wen} in the first loc: \textit{do moo\#| wiz\# and do \textit{is\#} wo dispijuut\#| wiz\# woz do \textit{strongo\#| wen o\# trava\#la\#| keem o\#}\#}.

\textbf{Note2:} syllable boundaries not yet shown: earlier rather than later, e.g. \textit{sok\#sil\#da\#| in \textit{mei\#ku\#}} do \textit{tra\#vo\#la\#} (except in words like \textit{maas\#to\#}, which go like \textit{aaf\#to\#}).

The timing shown was arrived at by applying the four ST rules. It seems a fair approximation. In this reading: five locutions or breath-groups, respectively 6 sec 6 moras, 6.11, 10.3, 4.8 and 4.10 long, the whole text adding up to 33 seconds 2 moras. Adding a couple of seconds for taking a breath between locutions, my reading of this text should take between 30 and 35 clock-seconds on the stopwatch for the occasion. It actually took 34.23 clock-seconds. Such `checking’ by watch or stopwatch may often be quite revealing for foreign learners as well. Later this year instrumental measurements/syntheses will be added to this paper to look at the syllable durations in more detail.

\begin{verbatim}
1 6 5 2 1 5 1 2 3 3 2 2 1 3 3 2 1 2 1 2 1 2 1 2 1 6 2 1 5 2 1 4 1 6 2 3 3 2 3 3 1 2 1 2 4 2 4 4 5 do moo\#| wiz\# and do \textit{is\#} wo dispijuut\#| wiz\# woz do \textit{strongo\#| wen o\# trava\#la\#| keem o\#}\#} 78=6.6

2 1 5 2 1 4 1 6 2 3 3 2 3 3 1 2 1 2 4 2 4 4 5 do \textit{ag\#\#la\#} do do \textit{waw\#| i\# \textit{faas\#}| sok\#sil\#da\#| in \textit{mei\#ku\#}} do \textit{tra\#vo\#la\#| bie\#k iz \textit{klo\#\#k \#}} 2 1 2 2 3 3 2 2 2 1 2 4 83=6.11

\textit{fog\# bi kon\#\#\#\#\#| strongo\#| don di mo#}

\end{verbatim}
Word-status: S, M, u, w.

**S-word.** Form: close juncture unit containing S-stress. Meaning: specified, ‘named’ or primary concept, requiring mental act of retrieval (= mental gesture generally accompanied by physical gesture).

**M-word.** Form: ditto containing M but no S-stress. Meaning: sub-specified, secondary or background concept (thought-in-thought); requiring ‘refinement’ of current retrieval act.

**u-word.** Form: ditto without S, M or *optional* w-stress. Meaning: sub-sub-specified concept = minimally specified concept = reference only to ‘automatically given’ concept; requiring automatic reflex only, no mental activity or gesture.

**w-word.** Form: ditto with *optional* w as highest degree of stress. Meaning: sub-sub-specified concept = super-ordained/preferred/highlighted reference or u-word, requiring minimal mental gesture. (Being relatively rare, w-word is here listed after u-word.)

[ø-word. In ellipsis the word/concept in question is completely unspecified.]

The following example (i) was already discussed in section 3. It is conceptualised as 1 locution or proposition, consisting of 1 piece or idea, consisting of 2 bytes or mental gestures (thoughts), consisting of 5 words or concepts, respectively uS|uuS. It takes 18 moras to say it, whether with pitch-jumps on hans and/or fa.o, or on a monotone.

\[
\begin{align*}
&\text{2} & \text{5} & \text{2} & \text{2} & \text{3} & \text{4} & \text{total 18=1.6 sec} \\
&\text{uS} & \text{uuS} & \text{u} & \text{S} & \text{u} & \text{S} & \\
\end{align*}
\]

(i) **jc æhays= /z \[n æfa.,#** 1 locution, 1 piece, 2 bytes, 5 words: uS|uuS#  

\[
\begin{align*}
&\text{5} & \text{6} & \text{4} & \text{2} & \text{3} & \text{4} & \text{total 24=2 sec} \\
&\text{S} & \text{S} & \text{S} & \text{u} & \text{S} & \text{u} & \text{S} & \\
\end{align*}
\]

(ii) **æjcc= æhays≠ æ.z= \[n æfa.,#** 1 loc, 2 pcs, 4 bytes, 5 wrds: S|S|S|S|S#  

\[
\begin{align*}
&\text{6} & \text{6} & \text{5} & \text{5} & \text{3} & \text{4} & \text{total 29=2.5 sec} \\
&\text{S} & \text{S} & \text{S} & \text{S} & \text{S} & \text{S} & \text{S} & \\
\end{align*}
\]

Example (ii) takes 6 moras longer to say than (i) because it is conceptualised/structured in a slightly different way. It is semantically structured as one locution consisting of two separate pieces or ideas. The first
The next two examples further illustrate these points. Compare the transcriptions (= ‘mental blueprints’), the timing patterns generated by applying ST1-4 to the transcriptions, the phonological structures in terms of the hierarchy locution>piece>byte>word>syllable, and the (semantic) word/concept ‘status’ derivable from that. Also note that these two examples introduce some more concepts: three M-stresses, but only two M-words, and two w-stresses but only one w-word. Finally note that example (v) consists of two locutions. This means they do not make up a single proposition, but two separate ones, the second one (with ø-specification of some concepts) only coming as an ordinary continuation or follow-up within the speaker’s larger conversational ‘turn’. [The extra time (usually) needed for taking breath was put in as an afterthought.]

Examples (iv) and (v) contain three M-stresses (with thanks to the author of the Longman Pronunciation Dictionary for pointing out an earlier error in the stress-marking of Eiffel Tower). Like everybody else, I sub-divide ‘stressed’ syllables into those taking primary or S(trong) and secondary or M(edium) stress.

The M stress in ætel/æv.xnn is a secondary or sub-stress within an S-word, and therefore does not affect the S-status of the word. It is rather like the secondary
beat on the third pulse in a four-to-the-bar musical rhythm. Musical rhythm is clearly hierarchical, in principle with an infinite number of groupings within groupings of one strong with one weak beat and/or one strong with two weaks. Such groupings, in other words, are just like the ‘feet’ employed in verse-rhythm. Tabla-players or jazz-drummers, for instance, build up extremely complicated patterns (and cross-patterns) ultimately analysable into hierarchies of iambs (ws), trochees (sw), dactyls (sww), amphibrachs (wsw) and/or anapaests (wws), interspersed with the occasional ‘mono-beat’ (s). Hierarchical organisation seems characteristic of all human perception and patterning.

In (iv), Ækw.in is given an M-stress, as is æafol in (v). So this makes them M-words rather than S-words. Not only do they take a sub-beat within a larger rhythmic grouping, they are also sub-specifications within a larger semantic grouping: Ækw.in is a sub-specification within the thought Æköw Æn æfgnol and æafol is a sub-specification within the byte Æz æt Æfæl æ'æpæz. Wheels within wheels, groupings within groupings, as usual. Here too, as so often in speech, one observes a direct relationship between form and meaning. In (v), on the other hand, the word Ækw.in is conceptualised as an S-word, specifying rather than sub-specifying the concept for the obvious reason that the Queen of France is not (usually) a single thought like the Queen of England.

Unlike many other phoneticians I also recognise a sub-division of so-called ‘unstressed’ syllables into those taking minimal or z(ero) and those taking non-minimal or w(eak)-stress. Assuming a ‘Rhythmic Alternation Principle’ (RAP) not allowing more than two ‘weaks’ between ‘strongs’, one simply has to do this if one does not want to end up with sequences of more than two ‘unstressed’ syllables. But I might point out that I recognised w-stress long before this argument occurred to me, simply on the basis of phonetic observation. So we arrive at a total of four degrees of ‘beat’ or stress: primary (S), secondary (M), tertiary (w) and quaternary (z), and a corresponding rhythmical hierarchy of foot within foot within foot. There seems no need (phonetically or semantically) to postulate more than four degrees of stress in speech, i.e. a rhythmical hierarchy of more than three layers. Cf. the far greater hierarchical depth common in musical rhythms, also assumed in the autosegmental-metrical view of speech-rhythm.

In example (iv) one must have a w-stress (phonetically very subtle and not immediately obvious) in the sequence of unstressed syllables gl,nd, shown by the small dot before the syllable in question. Moreover, one does this automatically and unthinkingly on unstressed syllables between two
unstressed syllables. Therefore it does not affect the u-status of the word \textit{iz} in the speaker’s mind. But in the sequence \textit{iz} \textit{at} \textit{di} \textit{ai} in example (v) one has a choice between saying \textit{iz} \textit{at} \textit{di} \textit{ai} and \textit{iz} \textit{at} \textit{di} \textit{ai}, the latter choice being less ‘automatic’ and therefore changing the word status from ‘given’ to ‘given but slightly preferred’. Note that w-stressing the weak form \textit{at} does not change it into the strong form \textit{æt}. Replacing the ‘empty’ vowel by a ‘full’ vowel would here immediately be felt as M or S stress.

Some may think that the distinctions just made are unreasonably subtle, indeed unnecessary. Subtle they may be, but if they are real, they are worth investigating. However, some foreign speakers of English have ‘unusual’ rhythms precisely because of unexpected distributions of w-stresses, so there is also a practical argument for describing it. The next section pursues this subject a little further.

9. w vs. z stress. ‘Refinement/reminder’ rule ST5: subtract point from z-values.

Words like \textit{beautiful}, \textit{political}, \textit{frighten}, \textit{comfort}, \textit{administer}, \textit{articulatory}, \textit{necessary} show how their w-stresses will ‘shift’ when one adds suffixes:  
\textit{bjouutt.fol}-\textit{bjouutfs.l}, \textit{politt.kol}-\textit{polittkks.l}, \textit{frætinn-}frætinng\textit{y}-frætinnng\textsl{y}.l, \textit{kamfs}-\textit{kamfs.tng}-\textit{kamfs.tngl}\textit{y}-k\textit{amfs.tngbol}.lns, \textit{admjnstr.tv}-\textit{admjnstr.tv.l}, \textit{ætjkj\textsl{lt}rt.l}, \textit{ætjkj\textsl{lt}rt.tl}, \textit{nessssslr}-\textit{nesssssl}l.\textsl{li} . Incorrect distribution of w-stresses by foreign learners may often be heard in such words with longish tails after the A\textsl{}ccent syllable.

It appears that the distribution of w in word-stress (but not in sentence-stress!) is completely predictable in English, and therefore need not be marked. This may be the reason why (unfortunately for the foreign student) pronouncing dictionaries completely ignore w-stress.

Reading out the following ‘realisations’ of \textit{how wonderful/comfortable that there/he was an attorney} with seven or eight ‘empty’ syllables in a row, may help to become aware of w-stress, and possibly of its phonetic exponents.

\begin{verbatim}
2- 3 1- 3- 2 1- 2 2- 1- 4 4- (Sequences of 7 \texttt{z}-syllables, some with w-stress)
(i) huw\texttt{wonderf}ol\textsl{y}.\texttt{d}ot \texttt{z} \texttt{wz} \texttt{en} \texttt{f3t3n}\textit{n}i \hspace{1cm} ‘neutral’ reading (hence no w-words?)
\texttt{u S u u u u S}
(ii) huw\texttt{wonderf}ol\textsl{y}.\texttt{d}ot \texttt{z} \texttt{wz} \texttt{en} \texttt{f3t3n}\textit{n}i\# \hspace{1cm} unusually (?) making \texttt{z} a w-word
\texttt{u S u w u u S}
\end{verbatim}
The phonetic differences between w and z stress, as said, are very subtle, but nevertheless quite audible (notably if consistently 'mixed up' by foreign learners). The most general feature, apart from feeling a difference in articulatory 'energy', is that z syllables tend to undergo even more shortening than w syllables, sometimes even to the point of complete elision. As a consequence of such further z-shortenings (indicated by minuses in the timing notations) example (v) is bound to sound slightly more staccato or 'syllable-timed' than (iv), which is indeed observably the case.

This could be accounted for thus: **ST 5. Substract another point from z syllables.** However, such a rule may not be too unpractical in computer-driven machine-synthesis, but it is for L2 learners. If properly applied they would have to subtract something from all z-syllable values, as we have only done in the present set of examples. Since w-syllables are very much rarer than z-syllables, it would obviously be much simpler to add a + to all w-values instead. But **Add another point to w-syllables** does not do justice to the phonetic facts. And so on. The cost-benefit ratio (a great deal of extra work to very little effect) being so unfavourable, it seems best to treat ST5 as a general reminder rather than apply it in all its detail.

10. S vs M vs z stress. ‘Refinement/reminder’ rule ST6: **Add 1 point to S syllable before M** (or: Add 1 point to syllable before sub-byte-boundary).
Apologies for the poor examples: *I like Jill Gascoyne, I can’t find a gas-coin (i.e. a coin for the gas), which phrases did Gus coin?*

Somewhat surprisingly, considering that the S/M=primary/secondary stress distinction is generally recognised, there seems to be no clear phonetic difference between them, even if speakers may perhaps feel a difference. If one says words like *underground, interchange, overall, anticlimax, reprimand, cigarette, mayonnaise, television, absolutely, catamaran, Waterloo, manageress, realise…* on a monotone (e.g. after the tonic shown in \( \text{a1} \cdot \text{dztnr} \cdot \text{zst} \cdot \ldots \) it is impossible to hear any difference between an S…M and an M…S pronunciation. So the S/M distinction seems to be purely a mental and structural one in the sense that only S can take a tonic, but not M.

The only reason for a special ST rule dealing with M stress is to account for the very slight (and not very crucial) lengthening of S followed immediately by M in the same byte. Rhythm being hierarchical the way it is, M-stressed or sub-stressed, sub-specified, \( \text{\&m} \) (in the example) is also a sub-grouping or sub-byte embedded in the larger byte \( \text{\&zts} \). So we may postulate a sub-byte-boundary after \( \text{\&zts} \). Since byte-boundaries require a slowing down or ‘ralentando’ by one mora (by ST3), it is not altogether surprising that the sub-byte-boundary here requires a lengthening or delay by less than a mora, say one point (indicated by a + in the timing notation). This suggests a rule ST6: Add 1 point to syllables before a sub-byte-boundary.

However, I fail to discern such a minimal delay and therefore indeed a sub-byte-boundary between non-contiguous M and S as in \( \text{\&k\text{t}am\text{\&}r\text{\&}m} \) - MzzS or \( \text{\&k\text{t}am\text{\&}r\text{\&}m} \) - SzzM, and similar words listed. Nor is there any lengthening of the M-syllable in MS compounds like \( \text{\&m} \cdot \text{kr\text{\&}m} \cdot \text{ice-cream} \) (vs. SM \( \text{\&m} \cdot \text{kr\text{\&}m} \)) Finally, examples of sub-byte-boundary placement contrasts in compounds like \( \text{\&m}\text{\text{\&}k} \cdot \text{\text{\&}k\text{\&}l\text{\&}v\text{\&}m\text{\&}k\text{\&}m} \cdot \text{make-believe} \) vs \( \text{\&m}\text{\text{\&}k} \cdot \text{\text{\&}k\text{\&}l\text{\&}v\text{\&}m\text{\&}k\text{\&}m} \), \( \text{\&m}\text{\text{\&}k} \cdot \text{\text{\&}k\text{\&}l\text{\&}v\text{\&}m\text{\&}k\text{\&}m} \cdot \text{ivy leaf} \) are difficult to find, but if they should have any rhythmic implication at all, it will also be marginal.

ST6 as just formulated again having such a poor cost-benefit ratio, indeed being partly incorrect, it seems best to formulate it for the time being as a simple reminder: *Add 1 point to S syllable before M.*

11. Given this analysis/description of English rhythm, the subsequent analysis/description of English intonation is extremely simple.

One can have rhythm without melody. One cannot have melody without rhythm. So before tackling English intonation, we should first sort out the rhythm, as I have just tried to do. It turned out that much of what is commonly regarded as ‘intonation’ is in fact ‘rhythm’. What remains will be
set out in my *Teaching the Rhythm and Intonation of English* to appear later this year on www.linguavox.nl