Rhythm in Speech and Language

A New Research Paradigm

Klaus J. Kohler
Institute of Phonetics and Digital Speech Processing
University of Kiel
Kiel, Germany

Short title: Rhythm and Prominence

Prof. Dr. Klaus J. Kohler
Holm 4
D-24113 Molfsee
+49 431 651453
kjk AT ipds DOT uni-kiel DOT de
Abstract

Like any other aspect of spoken language, rhythm needs to be, and has been, studied from four different perspectives for a comprehensive and insightful account of its nature and functioning in speech communication: symbolic representation, production, perception, communicative function. The paper first gives an overview of the milestones in the analysis of rhythm under the headings of these four approaches over the past 70 years. This survey of the development of scientific ideas in rhythm research prepares the ground for the integration of the four strands in an interrelated framework of linguistic and speech signal analysis. On the basis of a definition of rhythm derived from the theoretical and methodological discussion, a new paradigm is outlined for future research, centred on the listener and on communicative function.
1 Analytic approaches to speech rhythm

Like any other aspect of spoken language, rhythm needs to be, and has been, studied from four different perspectives for a comprehensive and insightful account of its nature and functioning in speech communication:

- **symbolic representation**
- **production**
- **perception**
- **communicative function**

The paper first gives an overview of the milestones in the analysis of rhythm under the headings of these four approaches over the past 70 years. This survey of the development of scientific ideas in rhythm research prepares the ground for the integration of the four strands in an interrelated framework of linguistic and speech signal analysis. On the basis of a definition of rhythm derived from the theoretical and methodological discussion, a new paradigm is outlined for future research, centred on the listener and on communicative function. The discussion is illustrated by audio examples and accompanying graphs, which can be found in a ppt file that is made available as an online supplement at http://www.karger.com/

1.1 Symbolic representation

This is the domain of linguistic categorization, which has dealt with two interrelated aspects of speech rhythm in intensive research over a long period of time:

- **accentuation** as a linguistic feature of making meaningful elements prominent
- **rhythmical timing** as a feature characterizing whole languages.

The discussion of **accentuation** has focussed, over many decades, on the analysis of languages that have **lexical stress**, especially Germanic languages, such as English or German, conflating this stress
level with the level of \textit{sentence accent}, and extrapolating phonetic features to other languages in a universal hierarchy of the acoustic properties 'fundamental frequency', 'duration', 'spectrum' and 'energy' for the signalling of stress, as initiated by the experimental investigations of Fry [1955, 1958] with synthetic speech. This has led to confusing terminologies and definitions (cf. Kohler 2008).

The concept of linguistic word stress has pervaded the discussion of rhythm for a long time. Anglophone writers since the 18th century have maintained, on an impressionistic basis, that in English stressed syllables tend to follow each other at isochronous intervals. Such a rhythmical characterization of a language on a stress principle was then expanded to a dichotomy of stress timing and syllable timing among the languages of the world, first proposed by Lloyd James [1940] with the terms 'morse-code rhythm' for a language like English and 'machine-gun rhythm' for, e.g., French. Pike [1945] coined the terms 'stress-timed' and 'syllable-timed', pointing out, however, that all languages contain both types of rhythmical structuring, but differ in favouring more the one or the other. This was later forgotten by other English-speaking scholars, who kept Pike's terminology but applied it as an either-or principle [Abercrombie 1967]. The rhythmic classification was eventually expanded to include mora-timing of Japanese as a third category, which was studied in a number of publications (see Warner and Arai [2001] for an overview).

To make progress in rhythm research in the languages of the world unequivocal definitions of accentual phenomena are needed. In the first instance, we have to distinguish accent categories at the lexical and phrasal levels, and may call the former \textit{lexical stress}, as part of the phonology of the word, and the latter \textit{sentence accent}, as part of the prosodic structure of utterances. In all languages, words may be represented by sequences of segmental vowel and consonant phonemes. In addition, languages differ in the way they make use of suprasegmental features at the word level. They may
use phonological pitch, in *tone* and *tonal accent* languages, or they may have *stress*, in stress languages, as a phonological place marker of a syllable of a word where a sentence accent manifests itself when the word is highlighted for propositional or expressive meaning. The place marker may be at any specific position in a word, with a high functional load for word differentiation, as in Russian, or it may be at a fixed position, with no functional load, as on the initial syllable in Finnish or on the penultimate in Polish, or it may be (in part) conditioned by morphology, with a low functional load, as in English or German; however, irrespective of the functional load, the position is essential for word intelligibility. *Stress* as such a phonological feature of the word is not defined by substantive parameters in itself, and it therefore does not make sense to describe it by values of acoustic or articulatory measurement, as has frequently been done, e.g. Fry [1955, 1958]. Such measurements of f0, duration, spectrum, energy refer to a sentence accent in a citation form utterance, at the place in the word that is phonologically marked for stress. Finally, there are languages, e.g. French, that do not make use of any suprasegmental feature at the level of word phonology, neither tone nor stress.

A definition of rhythm across languages cannot be based on a category of stress that is determined by lexical phonology. It cannot be based on the category of accent as defined above for the sentence level either, because the latter is linked to the highlighting of meaning. This is different from a regular recurrence of beats as the basic characteristic of rhythmic flow, which may be disturbed by propositional and expressive accentuation. The place markers for stress in languages such as English or German provide positions for rhythmic beats, but not all stress markers receive them, and, on the other hand, non-stress positions may get them. This was recognized in the study of Lea [1974] suggesting that whenever an increasing number of syllables intervenes between stresses, one of them will acquire the characteristics of a stressed syllable "to re-establish something like the ideal
alternation pattern". Various acoustic parameters carry these stress and non-stress beats: recurring f0 patterns, such as peak contours, and patterns of duration, energy, and spectral dynamics.

There is interaction of rhythmic structure with grammatical and pragmatic organization of an utterance, as was first acknowledged by Classe [1939]: "… in normal speech, sense is the chief thing, and other considerations will not prevail against it. On the other hand, our general speech habits tend to minimize the differences without quite obliterating them." [p. 78]. So, "three main elements [make or mar] the rhythm of an English sentence [as they coincide or work against each other]:

(a) The phonetic factor of number of units in the bar.
(b) The logical factor of grammatical connexion between the bars.
(c) The phonetic factor of the nature of the units in the bar, especially accents." [p. 89].

The rhythmic principle comes to the fore in idioms and popular phrases, such as the combinations, with and, of a monosyllable and a disyllable, which are felt to be more rhythmical than the reverse order because of the greater regularity of bars, e.g. English "with bow and arrow" and the German semantic reversal but rhythmical equivalent "mit Pfeil und Bogen". Jespersen [1905, p. pp. 220f] refers to this rhythmic principle in idiomatic phrases and lists many more examples of this type. Titles of books and films point in the same direction, e.g. Jane Austen's novels Pride and Prejudice and Sense and Sensibility, where the rhythmic beats are additionally heightened by alliteration. The German translations Stolz und Vorurteil and Gefühl und Verstand lose the alliteration but keep the rhythm, the latter with a different semantic ordering. The film version of the latter is marketed with the semantically wrong but rhythmically much closer title Sinn und Sinnlichkeit. In other cases, the ordering of binominals is clearly governed by meaning, the second of Classe's phrasing principles, for example father and son, mother and child, husband and wife, where the first element may be seen as
being closer to a prototypical orientation in society – adult, male [Lakoff and Johnson, 1980, Oakeshott-Taylor, 1984]. In sons and daughters, both principles coincide.

Lehiste investigated the interaction of syntax with rhythm in a number of papers, dealing, among others, with the disambiguation of sentences by duration (see [1980] for an overview). This is Classe's second factor of making or marring rhythm, listed above. If syntactic bracketing and rhythmic structuring interfere with each other, and if the speaker considers syntactic information important for the listener, the former will win against the latter, and the deviation from rhythmic regularity of duration makes the syntactic bracketing even more salient for the listener.

1.2 The speaker and the production of rhythm in the speech signal

When a concept of regular timing of rhythmic beats in either a stress-foot or a syllable frame had been formed as a characteristic feature of a language at the linguistic level of symbolic representation it was a natural further step to take it into the laboratory for phonetic substantiation through measurement of acoustic, articulatory, and physiological variables. The first studies analyzed isochrony of stress-timing in a language with lexical stress, viz. English. The general methodology of the early studies was the same: speakers of varieties of British and American English were recorded reading text, the investigator marked the stressed syllables perceived auditorily and then measured the durations from stressed syllable to stressed syllable in the acoustic or articulatory/physiological signal traces. This procedure is not without problems as regards the subjective assessment of stressed syllables, and where to start measuring – at the stressed-syllable onset (which may not be unique inside a word) or at the stressed-vowel onset or elsewhere.
Classe [1939], who was the first to tackle the question of foot durations in English, made kymograph recordings and thus segmented articulatory/physiological tracings. On the assumption that stress coincides with an increase of pressure in the speech-canal, he defined different measuring points for the segment types according to the mouth air trace: "(a) at the moment of release of plosion for breathed occlusives, (b) after the explosion of voiced occlusives, (c) at the moment of maximum deviation of the recording-pen for fricatives, (d) just before the beginning of the vowel in the case of all other consonants, except [h]". This anticipates the discussion of P-centres and their articulatory grounding 40 years later [Marcus 1981, Morton, Marcus and Frankish 1976, Fowler 1979, Tuller and Fowler 1980]. Classe failed to find isochrony, but given his theoretical system, in which rhythmic regularity may be made or overridden by three factors, allowed him to interpret his data as showing a strong tendency to isochronous feet, which surfaces provided certain conditions are fulfilled, namely a similarity of phonetic structure and number of syllables in the groups, similarity of grammatical structure of the groups, as well as similarity of connexion between the groups. These constraints guided his data selection for measurement and statistical analysis.

Subsequent studies of isochrony in the stress-timing of English analyzed the acoustic signal in the speech wave or spectrogram. Bolinger [1965], Uldall [1971, 1972], Lea [1974], measured durations of stress bars but without Classe's theoretical constraints. Therefore, a divergence from isochrony, which they found in their data, too, remained unexplainable and could thus only lead to the conclusion that rhythmic feet are not isochronous in English. Their attempts were doomed to failure because they did not determine thresholds for rhythmicity under various contextual phonetic and grammatical conditions. So, deviation from equal measures could neither support nor falsify the hypothesis of rhythmic isochrony as a feature of English. Moreover, these mechanical laboratory
analyses, treating speech as an object rather than as communicative behaviour, disregard the listener and the communicative function.

Lehiste [1973], however, based her data collection and analysis on Classe's conditions and recorded sentences containing four metric feet, which represented Abercrombie's [1964] mono- and disyllabic types, e.g. *Jack likes black dogs.* (monosyllabic), *Never visit busy cities.* (disyllabic short-long), *Jack has told me all about it.* (disyllabic long-short), *Always comfort needy orphans.* (disyllabic medium-medium). The average durations of the final metric foot were considerably longer than the other three, but remarkably similar, in all types, showing sentence-final lengthening. In any one type, the second and the third foot had very similar average durations. Both results confirm Classe's isochrony postulate under comparable conditions. The two centre feet show increasing durations from monosyllable to disyllable, and within the latter from short-long to long-short. Thus, bar durations are adjusted to meet the constraints of the syllables in them, but this does not invalidate isochrony as an underlying rhythmic principle in English (see also Lehiste [1980] for an overview of her work on isochrony).

Unfortunately, this line of research from Classe to Lehiste was not pursued further nor was it extended to other rhythmic language types. Dauer [1983] took up measuring inter-stress intervals again, without Classe's constraints, in a comparison of English, Greek, Italian, Spanish and Thai. Since duration measures were no more regular in a language, classified as stress-timed, like English, than in one listed as syllable-timed, like Spanish, she concluded that the search for phonetic correlates of the rhythm dichotomy could not be successful. Instead, she moved the research question back into the area of linguistic representation, maintaining that the rhythmic distinction between languages was not a phonetic feature in its own right, but a consequence of the differences in the languages'
phonologies, the most important of which relate to syllable structures, vowel reduction and word stress. Combinations of these phonological features create rhythmic diversity, which cannot form a dichotomy but must be continuous. This line of argument was also picked up in the publications by Bertinetto (e.g. [1989]). Finally, Nespor [1990] put forward the phonological view that rhythm was neither dichotomous nor continuous between stress- and syllable-timing, because there are rhythmically intermediate languages, such as stress-timed Polish with high syllable complexity but without vowel reduction, and syllable-timed Catalan with low syllable complexity and with vowel reduction.

The re-phonologizing of the research question has not advanced our understanding of rhythmic structures in the languages of the world, but engendered a new phonetic substantiation of phonological categories, with measuring procedures that gave up the search for isochrony of either interstress intervals or syllable durations. Ramus, Nespor and Mehler [1999] analyzed data from 8 languages (Catalan, Dutch, English, French, Italian, Japanese, Polish, Spanish) and computed three measures: %V the proportion of vocalic intervals within each sentence, ΔV the standard deviation of the duration of vocalic intervals within each sentence, ΔC the standard deviation of the duration of consonantal intervals within each sentence. These measures are located in a three-dimensional space, projected on the planes (%V, ΔC), (%V, ΔV) and (ΔV, ΔC) in Cartesian diagrams. They are assumed to be "an implementation of the phonological account of rhythm perception" in the 8 languages analyzed. The authors maintain that the (%V, ΔC) chart captures the variety of syllable structures because the number of consonants and their overall duration increase with the number of syllable types, resulting in a higher ΔC; since this also implies a higher consonant/vowel ratio, %V decreases. The positions of English, Dutch and Polish in the upper left-hand corner and of Japanese in the lower
right-hand corner of the (%V, ΔC) plane are regarded as an empirical validation of language rhythm contrasts on the basis of syllable structures.

The authors admit that the other planes are less transparent in their explanatory power of rhythmic properties. I would go even further and say that filling phonological structures with duration measures may be reflecting segmental duration patterns to a certain extent, but they are unable to capture rhythmic movement patterns evolving with a degree of regularity over time. Moreover, it is to be taken for granted that the analyzed data represent different degrees of rhythmicity due to individual variability on a scale from good to bad rhythm so that new data sets from the same languages will result in different locations of the languages in the three-dimensional space. The procedure is a form of data sorting but not an explanatory model of rhythm in speech and language.

The same evaluation applies to the measuring procedure proposed by Grabe and Low [2002]. They computed a vocalic and an intervocalic raw Pairwise Variability Index (rPVI) by first measuring the durations of vowels and the duration intervals between vowels (without pauses) separately in a passage of speech, then summing the absolute duration differences between two successive measurements, and averaging over the number of intervals. They also computed a normalized nPVI for vowel durations by dividing the absolute value of the difference between each pair of successive measurements by the mean duration of the pair. This was done to normalize for speech rate differences. The vocalic nPVI and the intervocalic rPVI for 18 languages (Catalan, Dutch, British English, Singapore English, Estonian, French, German, Greek, Japanese, Luxembourgish, Malay, Mandarin, Polish, Rumanian, Spanish, Thai, Tamil, Welsh) were entered into a Cartesian diagram to represent the rhythmic diversity of languages by acoustic duration measures. The authors interpret their results as supporting a weak categorical distinction between stress-timed and syllable-timed
languages according to phonological classification. There is considerable overlap between the two
groups and the hitherto unclassified languages, which could equally well point to a rhythmical
continuum. But, again, we are dealing with data sorting on the basis of consonantal and vocalic, i.e.
local segmental, durations in their point-to-point variability, not with global rhythmical patterns.

To advance the analysis of rhythm in language and speech we need to return to the line of research
intiated by Classe and followed up by Lehiste, and extend it to more variables than just the durations
of feet or syllables. What is essential for the production of rhythm in speech of whatever type and in
any language is the global temporal bracketing of the speech signal into chunks that have recurring
phonetic characteristics over and above the syntactic and semantic organization, but interacting with
it. These global variables are patterns of syllabic timing, of fundamental frequency and of energy,
recurrent with some degree of regularity over time. Especially the latter two define temporal chunks
that are long enough to allow for a good deal of variability, caused by more local segmental
constraints, without disrupting a percept of regularity for a listener.

But spectral patterning as a result of vocal tract dynamics is also a contributing factor to rhythmicity.
Therefore, we need to develop a model that can explain why *tit for tat, tick-tock, sing-song, ding-
dong, ping-pong, zigzag, flip-flop, wishy-washy* in English (Oakeshott-Taylor 1984) and *ticktack,
Singsang, dingdong, Pingpong, Zickzack, hickhack, klippklapp, Mischmasch, Wirrwarr, Krimskrams,
Wischiwaschi, Tingeltangel, lirumlurum* in German (Paul 1952, II, §131), with a close-open sequence
of vocal tract shaping, are more rhythmical than their reverse orders. The explanation may be sought
in the energy dispersion in the high-frequency spectrum for highish front vowels as against its
concentration in the lower-frequency spectrum for low and back vowels, coupled with higher vs.
lower intrinsic fundamental frequency. The sequencing of high-low spectral pitch may be perceived
as being more appropriately linked to the progression from the beginning to the end of a word or short phrase, just as, at the intonation level, high-low marks conclusion in phrasal utterances. Taking further into account that citation-form utterances of these words and short phrases in English or German get accentual patterns which favour high-low progression of fundamental frequency, there is coincidence of high-low spectral pitch and tonal pitch in one vowel sequencing but not the other. There is also the concomitant intrinsic shorter-longer duration, all converging on a perception of moving from the beginning to the end of a linguistic unit. This preference of vowel sequencing has nothing to do with lexical stress because some of the English examples have double, some single stress, nor is it causally linked to sentence accent, but it reflects rhythmic patterning for a listener that exploits the vocal tract acoustics of speech production.

We also need to explain why we not only associate Wolfgang Köhler's [1933] Maluma [ma'lu:ma] with a rounded figure and Takete [ta'ke:ta] with a zigzag one, but also prefer the ordering Maluma und Takete in the naming of the phenomenon. Continuous sonority is associated with the semantics of 'round, flowing, soft', voiceless obstruent interruptions of sonority with 'pointed, interrupted, hard', and we seem to prefer having the flowing section first, the interrupted one second, as a better rhythmic pattern. This is probably also the reason for preferring Gefühl und Verstand in the German translation of Sense and Sensibility to the semantically corresponding but reverse order.

Before physical measurement variables in speech production can be related to rhythmical patterns in a scientifically insightful way the type and degree of rhythmicity in the data needs to be evaluated perceptually by the competent language user. Native listeners have to scale how rhythmical native speakers' utterances are and assess whether some utterance is more rhythmical than another. It is only
then that acoustic or articulatory and physiological measures can be seen as the physical exponents of rhythmic categories in speech interaction in different languages.

1.3 The listener and the perception of rhythm in the speech signal

Since speech is produced for a hearer, signal parameters of rhythm, like any other speech property, are only relevant if they can be perceived and have a function for the hearer, who thus receives a key role in the manifarestion and analysis of rhythmic regularity. This was already recognized by Classe [1939], who distinguishes between measured objective and perceived subjective isochrony: "A certain irregularity of syllabic distribution will disturb 'objective' isochronism, but not necessarily 'subjective' isochronism," [p. 99] So, he maintains that the listener assists in establishing a concept of temporal regularity in speech by imposing a tendency towards isochrony on it. However, he did not run any experiments to substantiate his claim.

Isochrony in perception was investigated by Lehiste [1973, 1979]. She used a corpus of sentences consisting of four metric feet and supplemented them by replicas with noise-filled intervals separated by clicks for a perceptual test. Listeners decided in each case which of the four units was the longest or shortest. They had considerable difficulty identifying the longest or shortest measure in speech, but did much better on the non-speech material. The author argues that if listeners cannot tell the durations apart they must be perceived the same, so isochrony would be a perceptual psychophonic, as against a psychophysical, phenomenon, which, however, needs quantifying in terms of 'just noticeable differences' in reference stimuli.

She constructed further, noise-filled four-measure stimuli with three basic durations of 300, 400, 500 ms. Three of the four intervals were always the same For each reference duration, the length of the
remaining interval in 1st, 2nd, 3rd or 4th position was decreased and increased in nine 10-ms steps. Listeners were asked to identify the longest and the shortest interval, respectively, in two successive runs. The JNDs ranged between 30 and over 100ms, and they were bound to position in the stimuli: in the third interval, perception was most sensitive (30 to 40ms increment or decrement). Since most of the duration differences in the production of the four-metre speech stimuli were within this range of JNDs, and since the first perception experiment had established less sensitivity to differences in speech than in non-speech stimuli, Lehiste concludes that the deviations from physical isochrony in her production data are not perceptually relevant, and that isochrony can therefore be regarded as being (at least partly) a perceptual phenomenon. Donovan and Darwin [1979] confirmed Lehiste's results, adding that the perceptual phenomenon is not independent of pitch and that, as a feature of language, it reflects compensation for underlying processes in speech production.

These results point the way for future research into rhythmic patterning: to place the listener at the centre, even in the analysis of its production, in order to provide an insightful account of rhythm as a feature of speech interaction in languages. An essential part of this account is the active construction process by the listener, which Handel [1989, p. 449] describes as follows: "The acoustic wave induces us to hear a rhythmic pattern, but the acoustic wave does not directly signal that pattern."

1.4 The communicative functions carried by rhythm in the speech signal

The inclusion of the listener as a central authority in the analysis of speech rhythm also introduces the question of its communicative function. It has been claimed since antiquity that rhythm in sound generally and in speech in particular, has a guiding function for the listener. As Cicero puts it in De Oratore, Liber III, 185-186 (quoted according to the English translation in May and Wisse, 2001):
"…in all sounds and utterances rhythm is understood as the quality of having certain beats and of being measurable by regular intervals … if we are right to think that a constant steady flow of babble without pauses is crude and unpolished, the reason for this rejection is surely that it is natural for the human ear to measure the rhythm of the sounds that are produced by a voice, and that this is impossible if they don't have any rhythm; rhythm is the product of separation, of a beat at regular, or often varying intervals. We can discern it in falling drops of water (because they are separated by intervals), but not in an onrushing stream." And in comparing the orator and the untrained speaker, Cicero points out the contribution of speech rhythm to communicative success (ibid., 175-176):

"Among the many things that distinguish the orator from those unskilled and inexperienced in speaking, there is nothing that does so more than this: the unschooled speaker crudely pours out as much as he can, and lets his breath, not art, determine the limits of what he says. The orator, on the other hand, so ties his thoughts to the words, that all of them are encompassed by a kind of rhythm that is at once confined and free. For after fastening the thoughts in the bonds of form and cadence, he loosens and frees them by changing the order, so that the words are neither confined as if by some fixed law of verse, nor so free that they just wander about."

Classe noted the different rhythmicities of his subjects and took them into account in his data analysis, referring to a "rather poor rhythm" of his subjects 2 and 13 (op. cit. p. 52), and to subject 4 being "a very artistic speaker" who "has trained for the stage", so "we shall frequently use her recordings as a standard for comparison", (p. 63, fn. 2, p. 68). But in later analyses, by other researchers, of rhythmic measures in speech production, the rhythmic proficiency of individual speakers and the guiding function for the listener were on the whole ignored. Lehiste [1973, 1977, 1980] was the first to make a strong plea to incorporate function in the theory and empirical analysis of rhythm. She pointed out that in a language like English attention has to be focussed on stressed syllables because they carry
the greatest amount of information, and this is facilitated by rhythmic regularity creating expectancies as to when the next stressed syllable is likely to occur.

For instance, a sentence like To\textipa{d\textacircumflex{}\textquoteright{}ay the \textipa{w\textacute{}eather in \textipa{\textacircumflex{}Aix has a\textacircumflex{}\textacute{}\textipa{\textquoteright{}ain been} \textipa{\textacute{}abso}\textipa{l\textacute{}utely \textipa{\textacute{}w\textacute{}onder\textacute{}ful}}. can be produced with a perfect rhythm of waning trochaic prominence profiles in a sequence of foot structures that cut through words and do not just hook onto lexical stress positions but also mark other syllables, without putting any words in focus. This rhythmic sequence is more easily intelligible than the meaning-oriented structuring in Tod"ay the weather in Aix has again been "absolutely w"onderful. where foci of meaning are highlighted, and the prominences in between are suppressed, thus destroying the perfect rhythmic structure: the hearer is not guided through the sentence. So, regular timing contributes to optimal perception by the listener, making speech transmission and understanding faster and better.

Quené and Port [2005] turned this concept of a rhythmic guiding function into an experimental design measuring reaction times in a phoneme monitoring task. Starting from two target word lists of trochaic and iambic stress patterns in English, respectively, with one of the plosives /p, t, k, b, d, g/ in the onset of the stressed syllable (e.g. Pirate, campPaign), and another, very large set of trochees and iambs, they created sequences ranging between 5 and 7 words. Target words were in 5th position. The preceding and following words all had the same metric pattern, either trochaic or iambic, and the target word was either 'same-metre' or 'different-metre' (conditions of \textit{metrical expectancy}). For the alignment of the words in the listing sequences, a point was determined that marks the onset of the stressed vowel, using an adapted P-centre algorithm. Two alignment patterns were used, either regular or irregular spacing of alignment points in a sequence (conditions of \textit{timing regularity}). Each trochaic or iambic target word appeared in all 4 conditions. Subjects had to respond as quickly as
possible to the occurrence of a prespecified plosive by pressing a button. The reaction times were recorded and analyzed.

Reaction times to target words in same-metre and different-metre sequences were not significantly different; so *metrical expectancy* can be excluded as a factor in the processing of these listing sequences. However, *timing regularity* produced a highly significant effect, reaction times being about 60 ms longer in irregular timing. The effect is stronger for iambs than for trochees, but significant for both. The results show that *timing regularity* contributes significantly to spoken-word perception. Ease of perception depends on temporal alignment of the stressed vowels, irrespective of the metrical structure of the words. This indicates that lexical stress and, following from this category, the sequencing of stressed and unstressed syllables inside words is not relevant for rhythmic structuring and for facilitating word perception. Listeners' attention seems to be focussed on the timing of the rhythmic beat syllables. This supports Lehiste's view on isochrony as aiding speech perception. These results also fit in with the Dynamic Attending Theory (for a summary in relation to speech rhythm see Hawkins and Smith, 2001). In the regular-time condition, listeners' cycles of internal attentional rhythm were entrained to the external rhythm of the inter-stress intervals. So, listeners are able to focus on target words more quickly, yielding faster reaction times. In the irregular-time condition such entrainment to the external stimulus is not possible. Hence, listeners' internal pulsing tends to be out of phase with the most salient and informative parts of the external word sequencing, reducing the speed of perceiving the target words.

This is a promising line of research which needs to be put at the centre of future investigations into rythmic structures in the languages of the world. The quite limited paradigm of listing sequences will have to be transcended to include nursery rhymes, text reading, and finally spontaneous speech, with decreasing degrees of rhythmic regularity. But the incorporation of such other speaking styles needs
the initial analysis of rhythmicity in individual speakers, mapped on a scale from good to bad for the communicative function of guiding the listener. Listening to announcements in airports, planes, stations, and trains can convince anybody that intelligibility of spoken (in some cases vital) messages is severely hampered by speakers having poor proficiency in speech rhythm. Moreover, in spontaneous speech, rhythmic regularity is constantly disturbed by dysfluency, which may, however, be smoothed over perceptually by good rhythmicity in the adjacent fluent sections.

We must also be aware that there are additional factors apart from timing regularity that determine rhythmicity and subsequent ease of speech decoding. Advertising provides many examples of the intuitive knowledge of language users and speech designers as to what increases attentional focus in the listener. The case of marketing Jane Austen's novels and their film versions in English and German has already been mentioned (1.1). Another example is provided by the two English comedians John Eric Bartholomew and Ernest Wiseman. The latter changed his name to Ernie Wise early on. In their negotiations with television for a programme of double-act comedy shows it was suggested to Eric Bartholomew that he should also adopt a new name because "Bartholomew and Wise" did not have the right publicity appeal since it lacked rhythmic flow. When he said that he came from Morecambe in Lancashire this was picked up with "That's good. You are Morecambe from now on." So, in 1968, the BBC launched the very popular and highly successful "Morecambe and Wise Show". The two comedians had discussed, and settled on, this ordering of their names, ruling out the reverse because they considered it less flowing. The reason for this is no doubt the greater cohesion between the comedians' names and the name of the BBC programme if "Wise" comes second since unaccented "Show" does not get an additional prominence beat after accented "Wise", but would get one after the weak syllable "-cambe" of accented "Morecambe".
2 Parameters of rhythmicity

The preceding historical survey of rhythm research points to at least four parameters of rhythmicity in the analysis of production, perception, and communicative function. They are recurring timing patterns of

- fundamental frequency
- syllabic duration
- syllabic energy
- spectral dynamics

They are all produced and perceived as chunking of speech, by creating waxing and waning prominence profiles that occur with some degree of regularity over time. In this conceptualization of rhythm the focus is on global temporal patterns, as against a linear succession of local metrics, which are frequently restricted to segmental durations. F0 seems to be more powerful as a chunking property than the other three, although it has barely been included in the analysis of rhythm. It interacts with phrasal accentuation for propositional and expressive structuring of utterances, which may disturb the rhythmic flow. Syllabic duration and acoustic energy patterns also code global speech rate and loudness, and there is thus interaction between the rhythmic, rate and loudness levels.

Spectral dynamics seems to be the weakest of the rhythmic chunking parameters, as it is most closely related to much more local events. But it is subservient to the other parameters in 'stress-timed' languages like English and German, where function word reduction adjusts to the rhythmic exigencies created by the other three parameters.
For example, the German word "eigentlich" is generally used as a modal particle without information content, its main function being to soften the categoricalness of statements. In this function, it undergoes extreme, but systematic reduction (Kohler, 2001), e.g. [aɪn(ɻ)]. If there is accentuation for informational highlighting with the meaning "in reality", a much fuller form of the type [aɪŋθlɛ] is used to give the accented word more segmental body in addition to the prosodic features of accentuation, in accordance with the function of emphasis. The modal usage is illustrated by the utterance trifft sich doch eigentlich recht gut ("actually seems to fit quite well"); ['txif sɪç dɔç ʰaɪn ɣɛçɛ ɛ ɻuʈʰ] from the Kiel Corpus of Spontaneous Speech (Appointment Making Scenario, g092a021). But in spite of its segmental weakening, the word receives rhythmic prominence, just like "trifft" and "gut", resulting in a perfectly regular rhythm. On the one hand, this regularity is achieved by reducing the number of syllables of eigentlich from three to one, and concomitantly decreasing its consonantal complexity. On the other hand, a frame of waning prominence is set by a descending f0 scale, in parallel with an over-all energy decrease, across trifft sich doch. This is followed by an f0 rise out of low-frequency irregular vibration on [aɪŋ], and in turn, by a long stretch of voiceless friction, again of decreasing energy, with fricativization of the vowel in recht. The beat of this pattern on [aɪŋ] is strengthened by the glottalized onset of the word-initial vowel because word-initial glottalization features are tied to prominent and to accented syllables. The final plosive of recht has a very strong high spectral frequency fricative release; both the overall energy and the spectral frequency decrease in gut create a falling pitch and waning prominence perception. The result is a regular rhythmic timing: trifft sich doch | eigentlich recht | c gut, which, in cutting through the segmental structure of words, manifests its own organizational principle over and above lexical and semantic patterning. The falling spectral pitch and decreasing prominence in ['c gutʰ] also lead to the perception of an early peak pattern (Kohler 2006a), perceived as equivalent to a tonal pitch pattern.
The fundamental frequency and acoustic energy patterns are shown in Figure 1, and both graphic and audio illustrations are provided in Slide 1 under http://www.karger.com/kohler_phonetica-rhythm.ppt.

Although this is an isolated example, the account given here should not be regarded as an anecdotal description because it demonstrates two generalizable points:

- Rhythmic regularity is clearly perceivable, but it could not be captured by rhythm metrics such as PVI; it can, however, be mapped onto a bundle of physical parameters in the timing of fundamental frequency, overall energy and spectral dynamics.
- A fundamental distinction must be made between rhythmic prominence and informational or expressive accent (Kohler 2008).

It is a pressing demand on future research into speech rhythm to substantiate these observations by extensive experimental measurement.

Waxing and waning prominence profiles created by f0, duration and acoustic energy have long been known from psychoacoustic experiments with sine waves. Slide 2 http://www.karger.com/kohler_phonetica-rhythm.ppt exemplifies the grouping of tone sequences by loudness, timing and pitch in graphic representations and audio illustrations. Slides 3 and 4 provide corresponding examples of ba sequences grouped by f0 to single events, trochaic, dactylic, iambic, and anapaestic patterns (graphic representations of the speech wave timing and of three f0 patterns of 6 ba syllables in each slide with audio illustrations of 12 ba syllables). Slides 5 and 6 exemplify corresponding groupings to single events, trochaic and iambic patterns by syllabic duration and syllabic energy on flat f0 with graphic representations of the speech waves (6 ba syllables in the graphics, 12 in the audio). In each case, examples of simple ba syllable chains are turned into sequences of single events, or trochaic, dactylic, iambic, and anapaestic rhythms by overlays of f0, duration and acoustic energy patterns on otherwise
identical syllables. These waxing and waning prominence profiles in elementary articulated syllable
sequences recapture the groupings in sine waves. Slide 7 shows grouping by vocal tract shaping in
*biba* and *babi* syllable chains (3 *biba* sequences in speech wave, spectrogram, as well as f0 and energy
traces; 6 *biba* and *babi* sequences in the audio). The *bi* and *ba* sections are identical in f0 and in
duration and very similar in the energy traces. The audio series are clearly rhythmically structured by
the acoustic patterns of vocal tract shaping, but the rhythm may oscillate between trochaic and iambic.

It has also been known from psychoacoustic experiments with series of identical sinewave signals
that there is active construction: rhythm is in the mind of the listener (Handel, 1989). The same
happens with series of *ba* syllables. Slide 8 gives an auditory demonstration of the generation of
rhythm in psychoacoustic experiments with sinewaves, and in psychophonetic experiments with *ba*
syllables on level and falling pitch. The series of identical signals may be perceived as series of single
events or as trochaic, iambic, dactylic, anapaestic rhythmic structures if the listener wants to hear
them. The perception can be made to change in the course of a series.

As illustrated with the above example from German spontaneous speech, rhythm in speech is far
more complex because of the occurrence of syllable structures that can vary enormously in
complexity within and between languages. With the addition of different speaking styles of varying
degrees of rhythmicity from verse to reading texts to spontaneous speech there is an intricate network
of rhythm in the speech and language. This creates an enormously complex, but fascinating task for
the elucidation of speech rhythm in the languages of the world. From what precedes, the following
definition of speech rhythm may be proposed as a working hypothesis for upcoming research.
Constrained by the phonetic structures of the languages of the world, speech rhythm is the production, for a listener, of a regular recurrence of waning and waxing prominence profiles across syllable chains over time, with the communicative function of making speech understanding in various speaking styles more effective.

3 A new paradigm of rhythm research

This definition unites all four perspectives – the language, the speaker, the listener, and the communicative guiding function – for a comprehensive approach with listener and function at its centre. Based on it, a new paradigm of rhythm research is developed that includes the four temporal variables for the signalling of rhythmicity introduced in 2 – fundamental frequency, syllabic duration, syllabic energy, spectral dynamics – and incorporates them as acoustic parameters in a 3-part system of experimental design. Since prominence is a central concept in the definition of rhythm, it needs to be investigated at the outset. So, Part I of the new paradigm looks at the relative contributions of the four physical parameters to the perception of prominence and rhythmic patterning; Part II deals with rhythmical performance in scripted and unscripted speech by trained and untrained speakers, and with the perceptual evaluation of their speech productions; Part III brings in the factor 'language', applying the same experimental designs of Parts I and II to different languages.

3.1 Part I: The physics – perception link in prominence profiles

3.1.1 Step 1: The perception of prominent syllables

A first experimental analysis, using logatome baba bisyllables with German listeners, has been presented in Kohler (2008), showing that changes between first and second-syllable prominence can be effected by a systematic variation of f0, syllabic duration, and syllabic energy patterns across the two syllables, with f0 pattern as the most powerful. The procedure needs to be refined to a more
balanced design, and extended to more varied syllable structures, to more subjects, and to other languages. For details cf. op. cit.

3.1.2 Step 2: Scaling the contributions of the physical parameters to perceived prominence

The various physical parameters, f0, syllabic duration, syllabic energy and spectrum, need to be scaled in their individual and combined contributions to perceived prominence (cf. Kohler 2008).

3.1.3 Step 3: Generating rhythmical prominence profiles in syllable chains

The scaling of physical parameters in the perception of prominence patterns of disyllables can subsequently be used to generate trochaic and iambic rhythmical prominence profiles in multi-syllable chains and to successively expand them to bars containing more and varying numbers of syllables. These stimuli can then be evaluated perceptually as to their rhythmic regularity.

3.2 Part II: Rhythmical performance and its perceptual evaluation

The second part of the proposed paradigm investigates rhythmical performance by trained and untrained speakers in text reading, and its perceptual evaluation, testing the margins of regularity that have to be observed if speech is to be rhythmical. Standard texts representing different speech styles in a language will be used: (i) nursery rhymes, as elementary rhythms, (ii) verse, as highly stylized rhythm, (iii) prose text, e.g. The North Wind and the Sun, as less stylized rhythm. The aim is to obtain a spread of text productions from good to bad rhythmicity, which is in turn assessed by groups of listeners on a rhythmicity scale, e.g. on a 5-point-scale "excellent – good – satisfactory – mediocre – bad". The resulting data clusters are then analyzed for differences of syntagmatic structuring by f0, syllabic duration, and syllabic energy patterns, as well as their combinations, in relation to the results of Part I, to quantify the margins of rhythmical regularity. Slides 9 and 10 in
http://www.karger.com/kohler_phonetica-rhythm.ppt illustrate good and mediocre rhythmicity in English and German text reading, respectively, from the IviE Corpus and the Kiel Corpus of Read Speech (cf. also Kohler 2007).

3.3 Part III: The factor 'language'

The third part brings in the factor ‘language’. The same experimental designs of Part I and Part II are to be applied to languages that have different rhythmical structures. To start with, a Germanic language may be compared with French. The parameters of f0, syllabic timing, syllabic energy and spectral dynamics are expected to be combined differently to create rhythmicity in these languages.

In Germanic languages, rhythmic structuring cuts across syntagmatic phrasing by accentuation of lexically stressed syllables, as controlled by syntax and semantics. It is an interlevel between the accentual phrase and the syllable; prominence peaks in syllable chains for rhythmic patterning are typically, but neither exhaustively nor exclusively, associated with lexically stressed syllables. There is a tendency to compress syllables in between the rhythmic beats, but no isochrony of feet in the strict sense of the word, i.e. timing equivalence in observed measurements of speech production. Rather, rhythmical timing patterns are controlled by flexibility according to syllable structures within margins of regularity. This regularity needs to be mapped.

French, on the other hand, lacks this interlevel, having no lexical stress on which to hook recurring accents. Syntagmatic prosodic phrasing marks syntactic and semantic structures, and syllables enter the pitch phrases directly with their timing, energy, and spectral profiles. Pitch patterns that are controlled by syntactic chunking and semantic highlighting are not regularly recurrent and thus do not lend themselves to marking rhythmic units in speech. It is therefore to be expected that f0 is less,
syllabic duration more important in French than in Germanic languages to create rhythmical patterns. There is again no isochrony, but flexibility within margins of syllabic regularity.

This difference between, e.g., German and French comes out quite clearly in children's verse. Slide 11 [http://www.karger.com/kohler_phonetica-rhythm.ppt] provides audio illustrations.

*German*

Der 'Herr der 'schickt den 'Jockel 'aus,

er 'soll den 'Hafer 'schneiden.

Der 'Jockel 'schneidt den 'Hafer 'nicht

und 'kommt auch 'nicht nach 'Haus.

The rhythmic beats are marked by ' in front of the prominent syllables in the iambic bars. They create a succession of 4 or 3 waxing prominence profiles in each line although the beat syllables are not all equal in prominence. *aus* in the first line and *nicht* in the third are less prominent than the other 3 beats but more prominent than the second syllable of *Jockel* and *Hafer*. Similarly, in the second line, the beats in *Hafer, schneiden* and *soll* form a descending scale of prominence, but are, in each case, again more prominent than the preceding syllable. Thus, a regular rhythmic flow is produced across the whole verse.

*French*

Le bon roi Dagobert

Avait sa culotte à l’envers.

Le grand saint Eloi lui dit:

“O mon roi, Votre Majesté
Est mal culotté.”

“C’est vrai, lui dit le roi,
Je vais la remettre à l’endroit.”

There is no such regular recurrence of prominence patterns, organised in rhythmic bars, in the French verse, but a perceptual regularity in the sequencing of the individual syllables of each line, the ends of which are not only marked by rhyme but also by substantial syllabic lengthening.

The new paradigm proposed here for rhythm research should also be applied to languages in which pitch and duration are tied up at the word level and can thus no longer be used as freely at the level of rhythmic patterning, for example a tone language, such as Mandarin Chinese, or a language with quantity distinctions in lexically stressed and unstressed syllables, such as Estonian or Finnish or Czech.

4 Conclusion

The proposed new research paradigm takes variability in productive and perceptual performance as its point of departure to establish flexible rhythmic patterns. It assumes no surface isochrony, measurable in speech production, but gives the listener the key role in deciding on what constitutes rhythmic regularity. Rhythm of one form or another is considered scaled according to speaker, speaking situation and style and has a guiding function for the listener. Rhythm is thus not a fixed typological prominence pattern for groups of languages but is variable within each language.

However, it is also determined by the language in that the potential rhythmical parameters of f0, syllabic duration, energy, and spectral patterning over time are bundled differently in the languages of the world depending on how these parameters are tied at other linguistic levels. The analysis of structurally different languages in parallel according to the same experimental designs in production
and perception will be able to uncover the language-specific uses of the physical rhythmicity parameters in speech communication, and will thus provide a broader and more insightful basis for rhythm research in speech and language than the narrow focus on segment, syllable and foot durations of previous investigations.

Acknowledgement

Some of the ideas in this paper were first developed in a review of Andreas Dufter's PhD thesis. Parts were then presented at the annual meeting of the German Linguistic Society (DGfS) in Bamberg in February 2008, and at the Workshop on Empirical Approaches to Speech Rhythm at UC London in March 2008. The paper has profited from the dicussions at all these meetings, especially from Hugo Quené's references to reflections on rhythm since antiquity, as well as from Fred Cummins's assessment of André Classe's work: "his study is not given its due, although the relevant ideas for rhythm research are all there." I am particularly grateful to the two reviewers for the publication in Phonetica, Ilse Lehiste and Yi Xu, whose insightful comments have greatly contributed to clarifying a number of loose ends. Ilse Lehiste suggested that the experiment dealing with prominence perception (Kohler 2008), which was originally part of the present paper, should be published separately to make the rhythm paper more homogeneous. Remaining imperfections are of course my own.

References


Lloyd James, A.: Speech signals in telephony (London 1940).


trifft sich doch eigentlich recht [ʰ] gut

Fig.1. Speech wave, spectrogram, f0 (plain), and energy (dotted) of the German utterance *trifft sich doch eigentlich recht gut* from the Kiel Corpus of Spontaneous Speech (Appointment-Making Scenario g092a021). The three rhythmic bars have been marked by vertical lines across all three windows.