Domains of temporal control in speech and language From utterance to segment

Klaus J. Kohler

Institute of Phonetics and Digital Speech Processing (IPDS)
Christian-Albrechts-University, Kiel, Germany

ABSTRACT

With reference to Sagisaka's plenary paper the following issues of temporal control in speech and language are discussed: (1) the relationship between timing, duration and rhythm, (2) the distinction between units of duration and temporal effects of control parameters over varying domains, (3) the rhythm types of STRESS-TIMING, SYLLABLE-TIMING, MORA-TIMING, (4) timing in speech perception.

1 Introductory remarks

The Plenary Session paper by Yoshinori Sagisaka raises four general issues of temporal control in speech and language: (1) the relationship between timing, duration and rhythm, (2) the distinction between temporal units and temporal effects of control parameters over varying domains, (3) the rhythm types of STRESS-TIMING, SYLLABLE-TIMING, MORA-TIMING, (4) timing in speech perception.

2 Temporal categories in speech production research: TIMINIG, DURATION, RHYTHM

The temporal control of the vocal tract parameters, individually and in synchronization, their transformation into acoustic timing, and the mapping onto perceptual temporal patterns, are crucial for an analysis of speech communication. Therefore timing in the broadest sense, spanning all temporal aspects from production to perception, has occupied a central place in speech science research. It refers to speech parameters developing over time; it does not, per se, imply duration of specific units, such as phonetic segments or more abstract phonemes.

But this equation of timing with sequences of segmental/phonemic units, which are attributed durations, is what has dominated research into temporal aspects of speech. Sound segments are conceived as phonetic feature blocks or as templates with fixed multivariate parameter vectors, which have context-adjustable durations. The adjustment is achieved by time warp functions, which, in the simplest case of segment duration based synthesis, is uniform across the segment, with discontinuities at the segment boundaries. In a more sophisticated version of time warping of templates, the warp changes speed throughout the segment or in certain regions, e.g. the steady state [9]. But even in these cases the template vectors are not changed, on the assumption that asynchronous effects on parameter vectors are minimal and can be disregarded.

Modelling spontaneous speech has to cope with such asynchronicity. When a German speaker, under varying situational conditions, realises *wahrscheinlich (ein bisschen)* 'probably (a little)' as [vɑːˈʃamlɪç] or [vɑːˈʃamɪç] or [vɑːˈʃamç] or [vɑːˈʃaɪɲ], with successively greater compression, the first three pronunciations can be modelled by simply reducing some segments to 0 duration, but in the fourth one, the palatality of the whole third syllable, which is no longer manifested segmentally, is asynchronously hooked onto the nasal consonant of the second syllable. Since all these forms must be seen as related along a scale of temporal compression of articulatory movements, the time warping model cannot acount for all the facts. Such cases are very common in everyday unscripted interaction. We may ignore them and concentrate on the type of speech that is less likely to have such extreme timing contractions and that is more relevant in the practical application of speech synthesis, but then we bar ourselves from basic insights into how speech communication works: we have to acknowledge that segmental durations are an impoverished replica of speech timing.

Discussions of timing often refer to rhythm, either in an undefined, colloquial usage, or with the specific denotation of MORA, SYLLABLE or STRESS-TIMING. Under the latter perspective, the concept of duration units as temporal building blocks of speech is extended from the sound segment to hierarchically larger entities comprising the basic segment, i.e. to the mora, the syllable and the foot. Ever since Pike [8] introduced the dichotomy of SYLLABLE and STRESS-TIMED languages on an impressionistic basis, phoneticians have again and again attempted, through measurement, to establish the syllable or the foot (or the mora) as rhythmic units of duration above the segment in their own right.

ISBN 1-876346-48-5

In the original, extreme version of rhythm theory, the duration units above the segment were defined as isochronous, which means that the durations of segments inside morae and syllables, and of syllables within feet have to be adjusted to fit the constant timing pattern. The well-founded auditory impressions of rhythmic regularities, different for SYLLABLE, STRESS and MORA-TIMED languages such as French, English and Japanese were turned into tight duration measures, which in the case of stress-timing feet were modelled on musical bar notation, an abstraction from actual timing in performance. But two considerations preclude such an extrapolation from timing control to fixed duration a priori:

(1) With increasing complexity of segments in syllables and increasing number of syllables in feet, temporal compression of articulation reaches a limit, if all articulatory trajectories are carried out and the overall speech rate is kept. So there must be structurally conditioned variance of syllable and foot duration.
(2) There must also be random variance.

It is thus essential for the investigation into temporal control of speech to clearly differentiate the terms TIMING, DURATION and RHYTHM. The unfolding over time of physical parameters in speech production and their transformation into temporal patterns in speech perception under linguistic and communicative conditions is what we refer to as TIMING. DURATION is a constructive abstraction that is very useful, even if not sufficient, in connection with sound segments, but with higher units it has caused more confusion than helped our understanding [9, p. 239]. Finally, RHYTHM refers to the regular effects of temporal control over larger domains than the segment - mora, syllable, foot - without having to turn these domains into durational units.

3 A new look at the rhythmic typology of languages

3.1 Universal aspects of temporal control

Speech timing is controlled over intersecting, rather than strictly hierarchical domains of various extensions, ranging from utterance to segment. Some of these temporal control domains may be assumed to be universal:

Global utterance speech rate

It sets the frame for timing vocal tract trajectories and for pitch control over long stretches of utterance.

Temporal marking of phrase finality

Within utterances, syntactically and/or semantically/ pragmatically delimited phrases are usually marked by final lengthening. This is different from hesitation lengthening .

Vocal tract opening and closing gestures

Utterances are made up of temporal sequences of vocal tract opening and closing gestures with general articulatory constraints affecting their internal timing. It is in this gesture domain in particular that language or language group specific temporal control factors intervene, ranging from different rhythmical controls of gesture sequencing to timing of gesture components, under various prosodic (accent and intonation) as well as lexical and syntactic conditions.

Local emphasis

Gestures may be temporally intensified for semantic emphasis, especially their beginning, in certain positions of lexical items or syntactic constructions, e.g. *accent d'insistance* in French and non-pitch accents in German or English.

3.2 Pitch accent domain of timing

In languages that have pitch accents for putting words in focus and for marking their relative syntactic/semantic/pragmatic weight, e.g. the Germanic, the Slavonic, as well as Italian and Spanish among the Romance languages, sequences of pitch contours (peaks, valleys, hat patterns) can produce temporal regularity in addition to the regularity contained in the sequencing of opening and closing gestures of the vocal tract. The question then arises as to whether pitch timing control in these pitch accent languages defines a domain for independent, supplementary timing effects on vocal tract trajectories. For the Germanic languages the answer is "yes". Investigations have repeatedly pointed out temporal compression of accented syllables in relation to the number of subsequent unaccented ones, vowel reduction towards [ə], strong reduction of function words, levelling of unaccented opening-closing movements as effects of such an additional temporal control factor [2,3,5,7].

In spontaneous speech, unaccented syllable reduction between pitch accents can be very extreme, as in German *zwanzig Minuten* 'twenty minutes' ['tsvantsɪç mɪ'nu:tn] > ['tsvanʲ 'mnu:n], or in English *I don't know* ['aɪ doʊnt 'noʊ] > ['ə̃ə̃'ə̃] [4]. In both examples there are two rhythmic units in a two-pitch-accent pattern, which stay the same in the full and reduced forms. In the German example, only two accented syllables remain; in the English example, different degrees of opening and centralization of [ə]-type vowels are what is left to carry the accentual rhythmic structure. This extreme type of reduction can only occur in the right context of situation and in the proper prosodic frame: German *zwanzig* or *wahrscheinlich* will not be reduced to the same extent when they are not followed by another pitch accent to form a recurring rhythmic prosodic pattern. This pitch accent dependence together with a stripping to the bare bones of the accented opening and closing gestures within this rhythmic frame show that stress-timing is a separate factor. But this also underscores the necessity to incorporate segmental reduction in a model of speech timing, in a way that goes far beyond adaptive time warping.

ISBN 1-876346-48-5

In Italian and Spanish, on the other hand, there is no additional influence on vocal tract timing, as controlled by the pitch accent domain. In spite of pitch-accentual regularity, there is less compression and no vowel reduction [2,5]. Furthermore, these languages have quite simple syllable structures with a preponderance of open syllables [3]. Therefore syllabic regularity (not isochronicity) dominates the timing pattern and its perception, which has resulted in these languages being classified as syllable-timed.

3.3 Syllable-timed languages

French is different. Here pitch is primarily used as a phrasing marker, not for accentuation; focus, for example, is expressed by syntactic means. This entails that there is no pitch accentual regularity, intervening between the sequencing of opening and closing gestures and the utterance. Although syllables can be very complex in French (more so than in Italian or in Spanish) and heavy on consonants (e.g. *pas* ***de scr****upules* [d̥scr] or *parce que je crois* [$p^hsk^h\int kx^w\alpha$], where the voiceless plosive releases still have the function of syllabic nuclei), the general disregard for phonetic markers of word boundaries helps to create open syllables quite frequently. Moreover, syllabic reduction is only possible on condition that simple syllable structures result (cf. the rules of [ə] elision). So syllabic regularity (not isochronicity) dominates the timing pattern in French as well, but now in the absence of a superordinate regular pitch accent pattern. This in turn strengthens the perceptual salience of the regularity of vocal tract opening and closing gestures. Interspersing this syllabic stream with irregularly placed local emphasis markers linked to syllable-initial consonants (*accent d'insistance*) does not disrupt the rhythmic regularity, though it moves speech even further away from isochronicity. These features have created the picture o f French as the syllable-timed language par excellence.

Pike's bi-partition of languages was later expanded by the addition of the category of mora-timed languages, with special reference to Japanese, although there the mora as a timing concept is very old [10]. The extensive research of Sagisaka and his team, summarized in his plenary paper, focus on the mora as a unit of duration which has underlying isochronicity and whose modifications in the speech stream work towards maintaining a constant average timing unit. The main results are:

(i) A stronger negative duration correlation of a vowel with its preceding consonant than with its following one points to greater compensation within the CV mora, and this is an acoustic manifestation of mora-timing. Not all the work on mora-timing has provided the same results, and the methodological problem of using correlations between ajacent segments has also been pointed out [10]. More importantly, Kato's [6] perceptual findings argue in the opposite direction: the mora is not an auditory timing unit.

(ii) The greater the number of morae in a phrase, the shorter the average mora durations are. This is called local tempo, set by phrasing.
(iii) Overall tempo regulates mora-timing at the utterance level.
(iv) Final lengthening and initial shortening of single morae mark phrase boundaries.
(v) Boundary and local tempo effects on mora duration are independent of the word unit. Effects of function vs. content words are small.

This analysis of the mora as a unit of timing contrasts with its structural interpretation [1,10,11], which refers to the predominance of very simple (C)V syllable structures, to the lack of an accent effect on duration and to the irregular placing of pitch accents in syllable strings. All these factors guarantee a high regularity in the sequencing of CV morae by themselves, without the need for an independent temporal mora control. The lack of isochronicity is expected variance that does not interfere with this regularity, and compensation is therefore not necessary. Under this perspective Japanese turns out to be one of the most basic syllable-timed languages.

3.4 Segmental quantity contrasts

Abstractions from the opening and closing movements of the vocal tract lead to segments and their durations. The gestures of the vocal tract not only follow trajectories of widely varying complexity in different languages, they may also incorporate slowing down in the open or in the closed phase to implement phonological quantity oppositions between long and short vowels and between single and geminate consonants. In Japanese, this slowing down may either affct the V or the C of the basic opening gesture, and is traditionally counted as an additional mora each time.

3.5 Interaction of word structure with prosodic timing

Languages also differ as to the degree that word structure interferes with rhythmic structure. This interaction is minimal in French and Japanese, but is quite strong in Dutch, English or German. We may hypothesize that this interference can be greater in languages where pitch accent defines a timing domain, because a larger variance can be tolerated in the larger rhythmic grouping, without disrupting regularity.

3.6 Stylization

The choice of speaking style has two opposing effects on rhythmic timing. On the one hand, rhythmic patterns are more easily disrupted in spontaneous than in scripted speech because of the many additional interfering variables, from speech planning via word search to hesitation and truncation in discourse. So reading is more regular per se. But on the other hand, spontaneous speech shows extreme gestural reduction precisely within the typical rhythmical patterns. Finally, in nursery rhymes and poetry, we find a further

ISBN 1-876346-48-5

stylization of the characteristic rhythmic patterns of the language, strengthening their regularity. This resulted in the devices of syllable vs accent counting in Romance and Germanic metre, respectively.

4 Timing in perception

The temporal domain has received far less attention in speech perception than in speech production. Moreover, research into perception has been dominated by the segment and the phoneme [4]. This particularly applies to the investigation of rhythm. Kato's research [6] is one of the few investigations into the perception of mora timing in Japanese. By changing segmental durations in morae he produced larger perceptual effects on detectability and vulnerability for vowels than for consonants and showed a high correlation of judgements with the intrinsic loudness of segments. He also studied perceptual compensation between consecutive consonant and vowel durations by changing segmental durations separately or together in the same or in opposite directions. The drop of acceptability of these temporal distortions is greatest in uniform duration changes across both segments. This is interpreted as the lack of a perceptual compensation effect due to an inverse correlation with a jump in loudness at the segmental boundary. These experiments deliberately disturbed *perceptual coherence* [4] in the hypothesized temporal make-up of the mora time unit. The outcome is, however, not an experimental support for the mora as a perceptual unit with internal compensation, but for a psychoacoustic effect in relation to the construct of segmental duration.

What is needed is more extensive research into the modelling of rhythm perception quite generally within a theoretical frame that transcends the fixation on the segment and its duration. Dynamic Attending Theory seems to be a good candidate because (a) it rejects the representation of speech timing as time-warped sequences of underlying discrete elements, (b) replaces it by internal timescales attuning to those in external events through knowledge-driven expectations and focus of attention, and (c) postulates different time windows suited for different domains [4]. Thus if the English nursery rhyme of *Humpty Dumpty* is recited with typical French sounds and typical French rhythm, French listeners will perceive it as genuine, but meaningless French, whereas English listeners, familiar with the nursery rhyme, perceive a very heavy French accent on English words. If the French sounds are combined with a pitch accent rhythm, French listeners will perceive its non-Frenchness, and English listeners will probably recognise the nursery rhyme more easily. Here different timing windows and different expectations in these temporal domains together with differentiated focus of attention allow these complex judgements of the speech signal.

5 Concluding remarks

I have placed the work of Sagisaka and his group into the much wider field of basic research into speech timing in production and perception with special emphasis on different rhythmic structures in languages. I would like to stress in conclusion that their modelling of temporal characteristics of Japanese speech had its goal in engineering application to speech synthesis. In this scientific pursuit the fine detail of temporal control must be generated correctly for the synthetic speech output to be intelligible and to sound natural, and to achieve this aim the segment, its duration and its contextual time-warping are useful constructs. But the elucidation of speech timing needs to go further.

REFERENCES

[1] M. E. Beckman "Evidence for speech rhythms across languages", in *Speech Perception, Production and Linguistic Structure*, Y. Tohkura, E. Vatikiotis-Bateson, Y. Sagisaka (Eds.) , pp. 457-463. Tokyo: Ohmsha, 1992.

[2] P. M. Bertinetto. "Reflections on the dichotomy 'stress' vs. 'syllable-timing' ", *Rev. Phon. Appl.*, pp. 99-130, 1989.

[3] R. M. Dauer. "Stress-timing and syllable-timing reanalyzed", *J Phonet*, vol. 11, pp. 51-62, 1983.

[4] S. Hawkins and R. Smith. "Polysp: a polysystemic, phonetically-rich approach to speech understanding", *Italian Journal of Linguistics*, vol. 13(1), pp. 99-188, 2001.

[5] C. Hoequist. "Syllable duration in stress-, syllable- and mora-timed languages", *Phonetica*, vol. 40(3)), pp. 203-237, 1983.

[6] H. Kato. *Perceptual Characteristics of Temporal Structures in Speech: Towards Objective Assessment of Synthesis Rules*, Kobe University: PhD thesis, 1999.

[7] K. J. Kohler. "Invariance and variability in speech timing. From utterance to segment in German", in *Invariance and Variability in Speech Processes*, J. S. Perkell, D. H. Klatt (Eds.), pp. 268-289. Hillsdale: Lawrence Erlbaum, 1986.

[8] K. L. Pike. *The Intonation of American English*, Ann Arbor: U of Michigan Press, 1945.

[9] J. P. H. van Santen. "Segmental duration and speech timing", in *Computing Prosody*, Y. Sagisaka, N. Campbell, N. Higuchi (Eds.), pp. 225-249. New York: Springer, 1997.

[10] N. Warner and T. Arai. "Japanese mora-timing: A review", *Phonetica*, vol. 58(1-2), pp. 1-25, 2001.

[11] N. Warner and T. Arai. "The role of the mora in the timing of spontaneous Japanese speech", *J Acoust Soc Am*, vol. 109(3), pp. 1144-1156, 2001.