Chapter 4

Measuring Tempo

Introduction

4.1. Categorisation of tempo

Before talking about measuring speech tempo let us make clear with some examples what kind of tempo we deal with when we want to measure and to categorise tempo in speech. If we instruct two speakers to read a given text at three different speeds, first at a pace that is normal for them, then at a slow pace, and finally at a fast pace, then we have speech with three different tempi. If we measure the durations of each of these text recordings it can be assumed that the slow versions take longer than the normal ones, and the normal versions take longer than the fast ones. However, it might be that the slow version is shorter than the normal one, as happened with one speaker in Trouvain (1999). In other words, the durations of the various productions as *objective* measurements do not necessarily mirror the intended tempo as a *subjectively* produced speech tempo.

Moreover, if we ask listeners to judge which of two recordings of the same text they think is faster, the choice does not necessarily fall onto the production with the shorter duration. For this judgement, other factors could play a role such as dysfluencies, deletions and assimilations, number and duration of pauses (Goldman Eisler 1968) but there is also an influence of fundamental frequency on perceived rate (Kohler, 1986; Rietveld & Gussenhoven, 1987). That means that the *subjective* impression of speech tempo does not exactly match the *objective* measurement.

Thus, the following three types of speech tempo must be distinguished from each other:

- the subjective, intended tempo of speech production
- the objective, measured tempo reflected by durational correlates of linguistic units
- the subjective, perceived tempo

Tempo relates a distance to a duration, both measured with objective criteria. A tempo which is based on the relation of a "distance" in speech to a duration is able to quantify a given piece of speech *quantitatively on a continous scale*, e.g. in number of syllables per second. In contrast, the intended as well as the perceived tempo can be assigned to *categories*, e.g. slow or rapid. Each listener/speaker has an idea what slow, fast, normal (or however the category is named) means, but obviously everybody has her/his own interpretation of these categories, otherwise renditions of the same text at the same intended tempo would not diverge in their durations.

The intended as well as the perceived tempo compare speech tokens *relative* to one another. One can ask people to produce an utterance slower than normal, and people can judge whether a given utterance in one recording is faster than in another recording.

In order to make the tempo of instances of speech comparable, an objective metric seems the most promising method to do the job. The following sections deal with the problem of how to measure speech tempo quantitatively.

4.2. Units of tempo measurement

Measuring speed means relating a distance covered by a body to the time used. In speech, the articulators are our bodies moving in time and space. However, with respect to the articulators there is a lack of homogeneity: some articulators move inherently faster than others. The tip of the tongue e.g. is able to execute many more movements in a given time compared to the velum (Hudgins & Stetson, 1937, cited in Lehiste, 1970). Moreover, the articulators neither move all the time nor do the they move to the same extent. Measuring distances (here in millimetres) could of course be done for each of the different articulators. Although the generation of speech and its sound segments can be seen as the result of the execution of articulatory gestures (whose distance can be measured), it is crucial to understand that speech is the result

of the temporally coordinated execution of articulatory gestures that lead to speech *events*. That means, there is no distance that we can measure, we must seek a unit which describes speech events.

A number of different linguistic units have been proposed to serve as the substitute for a distance measurement unit in speech. In the literature we encounter a great variety of tempo denotations such as words per minute (wpm), syllables per minute (s/min), syllables per second (syll/sec or s/s), average syllable duration (ASD in msec), phones per second, or average phone duration (in ms). That means that units in use for measuring speech rate are, among others, the word, the syllable and the phone. Although commonly used, the definition of these linguistic units is not always straightforward. The advantages and disadvantages of these units shall be presented and discussed in the following sub-sections.

The word

Superficially, the word is easy to define and to count, and therefore apparently a useful unit for tempo measurement. A word can be defined as a sequence of letters that is not interrupted by a space or by an additional punctuation sign in a written text. However, the length of the units vary so much that the word is useless as a basis except for extremely long texts.

In some regards the graphical word is in conflict to other definitions of the word. The writing of the same words can differ within one language, depending on current orthography standards in a given language, e.g. in German "zusammenschreiben" vs. "zusammen schreiben"; English "shop-assistant" vs. "shop lifter", or "infra-red" vs. "infra red". Not every lexical word is expressed as one graphical word, e.g. French (and also German) "à la carte" or "San Francisco". Clitic groups such as German "ich hab's" or English "I don't" can be seen as two or three words. Also, the word can be seen as a morphological word, e.g. German "Berlin-Tegel" are two morpheme-based words but only one graphical word. Further, the number of words is unclear for many numerical expressions, e.g. German "17,50 €" are two graphical words, but in the spoken form three lexical words.

It would be reasonable for cross-linguistic studies to have a linguistic unit which allows comparisons across languages. Here, the length of words can vary to a very high degree, if we think of morphologically rich languages such as Finnish or German (e.g. German "Donaudampfschifffahrtskapitän") or agglutinating languages such as

Turkish. Similar lexical concepts should show a comparable length in compared languages. The three graphical words in the American English "Federal Supreme Court" are opposed to only one German "Bundesverfassungsgericht".

The syllable

A syllable can mean the underlying syllable derived from the lexical form of the word, or a syllable can mean the realised syllable. For the underlying syllable the number of syllables seems always clear. An exception to this clarity of syllable count is in German non-syllabic vowels as in *Piano, Lineal* or *genial*. In contrast to phonemic syllables the presence of a realised syllable is sometimes hard to detect. Syllables can be skipped (even words can be skipped or completely blended), and in many cases it is hard to decide when a syllable is skipped or still there. As a frequently occuring example in German, the phoneme sequence vowel-schwa-/n/ as in “ziehen” (Engl. "to pull") the syllable /ən/ can be realised as a syllabic [n̩] or as an non-syllabic [n], leading to different syllable counts.

The sound segment

One interpretation of a phone is the phonemic segment of a lexical word, whereby the phonemic status of certain sound segments are still a matter of debate. Another issue is whether affricates and diphthongs should be seen as mono- or bi-phonematic, i.e. one or two segments. Usually, a glottal stop is not given a phonemic status. Further delicate aspects include the results of phonological processes such as diphthongisations (e.g. in German homosyllabic vowel+/r/-sequences like in "Start" (Engl. "start"); or the phonemics of certain affixes, e.g. the "er" in "ersetzen" (Engl. "to replace"), which could be /ɛr/, /ər/, /ɛɐ/ or simply /ɐ/; or the degemination of homorganic consonants as in "kann nicht" (Engl. "cannot"). Even if the listed problems do not contain central concerns about the sound segment as the optimal unit for tempo measurement it should be clear that the segment is not unproblematical.

Intended vs. realised forms

A distinction which is infrequently made is between the intended forms (corresponding to the canonical or lexical or underlying form) and the realised forms, the latter is called "effektive Lautzahl" (Engl. "effective number of sounds") by Hildebrand (1963). *Intended* forms have the advantage that they can be easily derived from the lexical representation of the uttered words, whereas their actual *realisation* can vary strongly. This fact has already been pointed out by von Essen (1979) and can be illustrated with the German sentence "Am blauen Himmel ziehen die Wolken." (Engl. lit. "In the blue sky wander the clouds."). The transcription of this sentence consists of 26 phonemes and 10 syllables:

/ʔam blaʊən hɪməl tsiːən diː vɔlkən/

However, a typical reduced realisation of this sentence, is shrunk to 20 phones in 8 syllables:

[am bɭaʊn̩ hɪm̩l tsini vɔlkŋ̍]

If we assume a duration of 2 seconds for a realisation of this sentence, the measured or "objective" tempo in phones/sec would either be 13 phones/sec (intended) or 10 phones/sec (realised); the "objective" tempo in syllables/sec would either be 5 syll/sec (intended) or 4 syll/sec (realised). Ironically speaking, a speaker can speed up or slow down the speech tempo by a quarter just by defining the unit of measurement. This example shows that just one criterion of the definition of the unit of tempo measurement, here the sound segment, can be decisive on the meaning of what has been measured. This, of course, has serious implications for comparing data of different studies.

Other units

In music, tempo is measured by a metronome in beats per minute. The composer can either indicate the metronome value or can use a tempo term such as *adagio, lento, largo, grave* for slow tempi and *moderato, allegro, vivace* for faster tempi. These terms correspond to metronome values where the normal range is considered to lie between 75 and 80 beats per minute, i.e. values which are slightly higher than the 72 heart beats per minute of a middle-aged healthy adult person.

The idea of also using beats per minute in speech has been applied by a few researchers, e.g. by Uhmann (1989). In addition to syllables per second, she proposes *accents per second* as an additional measurement of tempo or "density". In her analysis of German conversational data she has shown examples of what she calls "contextualisation cues" in which speakers make utterances interpretable in dialogues. For example, a low number of accents per second combined with a high number of syllables per second serves to contextualise parenthetical utterances, sidesequences and afterthoughts. In contrast to these passages of low relevance, portions of high relevance such as emphatic utterances are contextualised by a high number of acc/sec and a low number of syll/sec. A combination of a high number of acc/sec and a high number of syll/sec can be found in repair sequences. The problem with accents is of course to define this unit with the aim of a consistent and reliable use across researchers. Uhmann (1989) transcribed primary, secondary and emphatic accents, but the transcription of these can differ between labellers, which is counter-intuitive to the idea of having a quasi-exact quantification.

Last but not least, non-linguistic units that were derived from the acoustic signal have been applied to quantify speech tempo. In studies aiming at detecting articulation rate automatically, e.g. for use in automatic speech recognition. This is done to improve the modelling of fast speech with a high number of segment deletions and replacements (cf. chapter 3) usually featuring an disproportionally high word error rate. Morgan, Fosler & Mirghafori (1997) calculated energy fluctuations to determine articulation rate whereas Samudravijaya, Singh & Rao (1998) enhanced the parameter set and also tested measures of non-stationarity and voicing switch rate.

Selecting a unit of tempo measurement

The previous sub-sections make it clear why there cannot be an objective "metre" for speech tempo measurement. Nevertheless, one linguistic unit must be selected if speech tempo is to be quantified. The following criteria may give an orientation for selection:

- degree of popularity
- comparability across studies
- ease of counting
- ease of definition

- reflection of temporal variance

The word (as words per minute) and the syllable (as syllables per second or as average syllable duration) seem to be widely used as tempo metrics, whereas the sound segment (usually as phones per second) seem to be less frequently used. Regarding the comparability to data of other studies of the same as well as of a different language, the syllable and the sound segment seem to be preferred rather than the word. Counting tokens is not a problem for the word, and counting does not cause greater problems for the syllable. However, counting sound segments requires a transcription of all recordings, and that is often not available. The easiest definition can be given to the phonemic syllable followed by the word, and here again the sound segment seems to be the most problematical case. Nevertheless, the essential characteristic of a unit expressing tempo is the reflection of temporal variance. Here, the word seems to score worst, and the sound segment best, followed by the syllable, i.e. the smaller the better.

In order to check the tempo fluctuations due to the choice of the unit, Carroll (1966, cited in Kowal, 1991) investigated which differed in the number of syllables per word in a reading aloud experiment texts. The measurement of words, syllables and phonemes per minute showed that the variation coefficient of the mean values per text was highest for the word and lowest for the phoneme. The most reliable results for the different texts were found for phonemes per minute.

This finding is in agreement with the results in Trouvain et al. (2001) with German data where the number different linguistic units were correlated with articulation time. It was shown that realised phones correlated best, followed by intended phones, realised syllables, intended syllables, and words (in this order).

In a study recommending standard speech rates for foreign language training, Tauroza & Allison (1990) compared the word rates and the syllable rates of four different speech styles. For reasons of different word-to-syllable relation for each style (news texts having more syllables per word than interview speech), the two rates were not at all in agreement with each other. Syllable rate was found to be better as an expression of one standard tempo for various styles than word rate.

An argument against the syllable as a quasi-universal unit is that in mora-timed languages such as Japanese, speech tempo is frequently measured in mora per second (e.g. Kuwabara, 1996; Koiso, Shimojima & Katagiri, 1998).

Also in testing the sensitivity of different tempo measurements for the classification of (English) speech according to their tempo (for use in automatic speech recog-

nition), the recognition rate is more sensitive to phone rate than to word rate (Siegler & Stern, 1995).

Assessing the contributions of words, syllables and segments to utterance duration with reference to articulation rate measurement, Faulkner (1997) identified for English texts the phoneme as the single most significant variable to explain durational variance.

Den Os (1985) gives evidence that phonological syllables per second and phonetic segments per second best fits the perceived speech rate for Dutch and Italian short utterances. Phonetic syllables were worst.

In perception tests investigating the estimation of local speech rate, Pfitzinger (1999) found out that a combination of phone rate and syllable rate matches the subjective evaluations best when listening to short windows of speech (625 ms).

Referring to differences between languages, where the rhythm type of the language play an important role, Roach (1998) favours the sound segment as unit to be preferred:

> "Dauer (personal communication) has found that Greek and Italian are spoken more rapidly than English in terms of syll/sec, but this difference disappears when sounds/sec are counted. […] It seems that on the evidence available at present, there is no real difference between different languages in terms of sounds per second in normal speaking cycles."

To summarise, among the existing units there is obvioulsy not *the* optimal unit for tempo measurement. The selection of the unit depends on the purpose of the study. However, although word per minutes seem a rather widely used metrics it is obviously less favourable for most purposes. An exception may be when more abstract units are compared, as done in the study by Grosjean (1979) who investigated the articulation rate and the pause rate of signers (American Sign Language) and speakers (American English).

The criteria listed have not been weighted so far, but it seems clear that the unit that mirrors tempo best is the one that is most sensitive to temporal variance. By nature this is the smallest unit, i.e. from the units presented here the (realised) sound segment. However, there are other factors worth consideration. One usual way to

economise articulatory effort, with the consequence of speaking faster, is to reduce the number of realised segments. That means that the intended phone would be the appropriate candidate because it additionally accounts for an important tempo variation factor, anemely degree of segmental reduction. And the last note on the relative importance of the listed criteria refers to the ease of definition and the ease of counting, which speaks for the intended syllable (ignoring the word). These two criteria will be the decisive ones for many studies and many applications, simply for practical reasons.

The role of pauses in tempo measurement

With reference to the beginning of chapter 2, articulation rate was defined as the net speech rate, and speaking rate including the pauses was defined as the gross speech rate, in line with many other researchers (e.g. Goldman Eisler, 1968; Wood, 1973). A look at table 2.1 (p. 7-8) makes it clear that the differences can be substantial between these two measurements, ranging up to several syllables per second difference for the same recordings, especially in spontaneous speech with a high percentage of pausing time.

If the differences can be so dramatic, then the defintion of a pause is crucial to determine speech tempo. In chapter 3, several thresholds in different studies were listed, ranging from 50 ms up to several hundred ms. It goes without saying that an articulation rate measured with a pause threshold of 100 ms can differ considerably from the articulation rate measurement of the same recording with a pause threshold of 500 ms (cf. Kowal, Wiese & O'Connell, 1983).

Besides pause thresholds, unintentionally articulated speech also causes problems, e.g. in a filled pause ("die äh meiner Meinung nach") or in corrections of slips of the tongue ("die deiner Mei, nein meiner Meinung nach") or in word repetitions in spontaneous speech ("also die die die nicht das nötige Kleingeld haben"). There is no common standard whether to consider these dysfluencies as ordinary speech articulation, or as a pause, or simply to ignore these instances of badly formed articulatory performance.

4.3. Dynamics of global and local tempo

Global and local levels of articulation rate

Another uncertainty when dealing with speech rate concerns the stretch of speech taken into consideration. Speech rate changes continuously while speaking (cf. Wood, 1973; Miller, Grosjean & Lomanto, 1984), so that the first part of an utterance can be spoken fast, while the second part can be rather slow, or vice versa. An average rate calculated for an utterance does not necessarily reflect the tempo characteristics of different parts. When the domain is not specified, it is not clear whether the speech rate quantifications are related to a more global or to a more local level. Most of the time, when people talk about speech rate, they use the term globally, referring to an entire text, sentence or whatever the utterance might be. The problem of local variations has long been neglected. The main question to be answered is: How "locally" should speech tempo be considered?

No matter what the local unit will be, despite one global rate that can be determined, there are tempo differences between the individual phrases. Spontaneous speech can be expected to be marked by more changes in articulation rate than we find in read speech: Planning problems are likely to cause hesitations (e.g. syllable drawls) leading to slow stretches followed by fluent, fast stretches. These planning problems in spontaneous speech also increase the number of filled and unfilled pauses which lead to shorter inter-pause stretches. Especially utterances consisting of only one or two discourse particles such as "ja" contribute to a high number of short but very slow inter-pause stretches. The last points would support the reported tendency that "the longer the utterance the faster its rate" (cf. Fónagy & Magdics, 1960; Malécot, Johnston & Kizziar, 1972, Martínez et al., 1997, but see also Koopmans-Van Beinum & Van Donzel (1996) for different results). Emphasis, which occurs more often in spontaneous speech, represents another factor which results in a slower tempo.

In an inspection of the German "Kiel Corpus of Read and Spontaneous Speech" (IPDS, 1994) we compared the rate characteristics of read versus spontaneous speech (Trouvain et al., 2001). The results of this study (replicated in table 4.1) show that in spontaneous speech inter-pause stretches (ips) as well as intonation phrases (IP) are shorter on average and show a greater variance than in read speech.

Table 4.1: Mean duration (in sec) and mean articulation rate (real. phones/sec) of inter-pause stretches (ips) and intonation phrases (IP) for spontaneous and read speech with standard deviations.

		duration mean (sd)	**articulation rate mean (sd)**
spontaneous	**ips**	1.81 (1.29)	13.24 (3.29)
	IP	1.17 (0.73)	13.18 (3.75)
read	**ips**	1.98 (1.03)	13.06 (2.03)
	IP	1.49 (0.67)	13.01 (2.23)

With respect to articulation rate, spontaneous speech is slightly faster and shows a greater variance (see also figure 4.1). Although faster on average, spontaneous speech features a high number of slow utterances. One reason lies in the large number of very short inter-pause stretches (<1 sec) in this speaking mode. Indeed, one and two word utterances are slower than the mean. Intonation phrases are generally shorter than inter-pause stretches, but there is basically no difference in articulation rate.

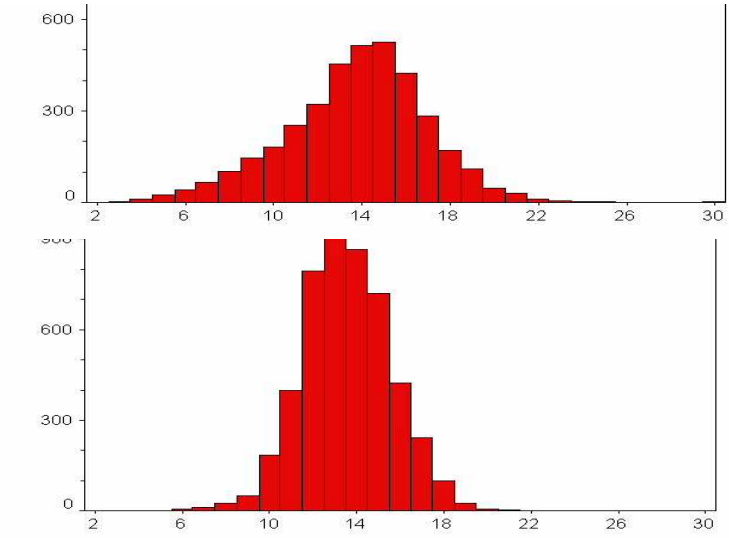

Figure 4.2: Histograms of articulation rate (realised phones/sec per inter-pause stretch) in spontaneous (top) vs. read speech (bottom) in the "Kiel Corpus of Read and Spontaneous Speech" (data from Trouvain et al., 2001).

Domains of articulation rate

When we looked for the "optimal" unit to describe tempo, the main criterion was that this unit expresses the temporal variability best. Now, searching for the "optimal" domain, we look for a stretch of speech in which the tempo variation is smallest, or, in other words, where articulation rate shows the highest degree of constancy.

Whatever the optimal utterance domain may be, tempo changes can occur not only between adjacent phrases but also within a phrase. The problem lies in the acceleration and deceleration within the local section. Each syllable lengthened due to accentedness or phrase finality is decelerated. We can focus domains as small as the syllable or even the syllable rhyme (phrase-final lengthening). All these very local phenomena can be seen as *accelerando* and *rallentando* (or *ritardando)* as labelled by Crystal (1969) in his list of prosodic systems under the heading *complex tempo system* in addition to the *simple tempo system* of global rates such as allegro and lento.

The previous paragraphs showed that the global tempo of a longer utterance can be distinguished from the local tempo of a single phrase within this utterance (confer left and mid pattern in figure 4.1). Moreover, there can be tempo variations within this single utterance showing e.g. a rallentando pattern, as illustrated in figure 4.2 (right side).

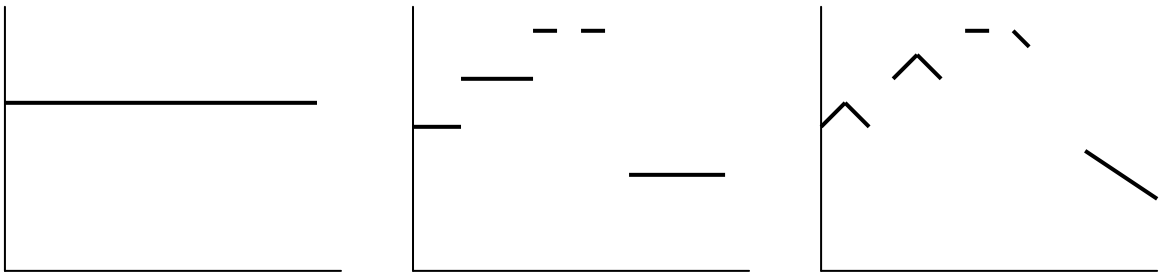

Figure 4.1: Global and local levels of tempo in idealised schemata of time course (x-axis) and tempo (y-axis); left: global rate for the entire utterance; mid: local rates for single phrases (e.g. inter-pause stretches); right: local rate shapes within the single phrases.

In her studies of Czech and British English, Dankovičová (1997) investigated the following spans of speech production as domains to measure articulation rate: the in-

ter-pause stretch, the intonation phrase and the syntactic clause. She showed that the duration of phonological words is best mirrored by the intonation phrase.

Normalisation of rate dynamics

Crystal & House (1990) showed with articulation rates in a reading task with American English speakers that slow talkers and fast talkers differ in their global tempo (as illustrated in figure 4.2 left side) and they differ in the rates of the inter-pause stretches (cf. figure 4.2 mid). However, slow and fast readers have very similar patterns of local rate changes, i.e. the pattern in figure 4.2 (mid) is shifted upwards for fast speakers and shifted downwards for slow speakers.

The aim is to determine and to weight the factors responsible for the variation. Such a normalised value makes it easier to compare utterances differing in rate. But how can we relate given (phonological) information about syllable structure, number of segments, phrasal stresses, phrase boundaries and so on to a "normalised" rate? Koopmans-Van Beinum & Van Donzel (1996) tried to do so by assigning different weights to various phonological factors such as vowel quantity and schwa syllable. Although they consider their attempt to normalise rate dynamics in inter-pause stretches as preliminary they were able to show that the normalised rates of these phrases point to the discourse structure of the text. Phrases which are used to introduce something new are marked by a slow normalised rate. This picture was not so clear without the normalisation. It might be that such a normalisation of speech tempo could be a helpful instrument in order to improve the detection of temporally marked elements of information structure. This was also done by Uhmann (1989) who identified in her data less relevant passages such as parentheses with a high syllabic rate and a low rate of pitch accents and, in contrast to that, highly relevant passages with a low syllabic rate and a high density of pitch accents (cf. also Barden, 1991). However, looking at the experiences of Koopmans-Van Beinum & Van Donzel (1996) there is a big need for research:

> "The main conclusion of our study must be that accounting for variations in speaking rate of what may be considered as 'spontaneous speech', is a very complicated task."

It is one thing to normalise objective tempo by calculation, it is another thing to test how actual listeners normalise for, or indeed whether or how they perceive tempo

variation found in objective tempo values. A listener appears to compensate for the numerical variation in rate, a fact that can be explained by the linguistic and phonetic (rhythmic) restrictions. There are comparably few instances of noticeable tempo changes in spontaneous speech (Batliner et al., 1997) and there are expected to be no noticeable tempo changes in neutral read speech, e.g. news reading.

Summary and discussion of chapter 4

Measuring speech tempo contains various sources of confusion. In this chapter we attempted to make clear that we must distinguish whether tempo means the intended tempo category in speech production, or a perceived tempo category, or a quantified objective tempo, where acoustic correlates of linguistic units are related to physical time.

The latter consideration is often expressed as word rate, syllable rate or phone rate, leading to the central question of speech tempo measurement: what is an optimal unit to quanitify speech tempo? What are criteria to determine an appropriate unit? The pros and cons of the syllable on the one hand, and the sound segment on the other were discussed. The word was considered to be the least optimal tempo unit – despite its frequent use, e.g. in speech synthesis markup languages such as SABLE (Sable URL). Although there are many arguments for the sound segment as the preferrable unit, the arguments from a practical perspective favour the phonological syllable as standard unit for tempo measurement.

But apart from the unit itself, further perspectives should be taken into consideration when speech tempo is quantified. These include the vital role of pauses (leading to a net rate excluding pauses or a gross rate including pauses), consequently the definition of a pause (there are great variety of thresholds), and also the domain in which articulation rate is measured.

The last sub-section was dedicated to the question of tempo variability found across and within phrases, with supporting data from our own corpus analyses. This led to a further distinction to bear in mind, namely the necessity to keep apart the global tempo of a longer utterance from the local tempo of single phrases within this utterance.