The rhythm window

OLIVER NIEBUHR
Doctor in Linguistics. Dept. of General Linguistics,
ISFAS, Universityof Kiel, Germany.
niebuhr@isfas.uni-kiel.de

Resumo: O artigo apresenta os resultados de uma pesquisa combinada sobre produção e percepção do ritmo na fala do alemão. Os resultados relativos à percepção demostram que só é possível uma identificação de ritmos complexos em uma velocidade de fala (proporção silábica) de 4 a 8 sílabas por segundo. Mesmo estímulos acusticamente monótonos causam „ritmos subjetivos" nessa gama de valores. Em contrapartida, a percepção do ritmo fora dessa gama de valores fica aplanada. Os dados de produção concordam com tais resultados. As velocidades da fala cotidiana variam entre 4 a 8 sílabas por segundo e não descem abaixo desses valores, a não ser que os falantes aplanem seus ritmos para dar ênfase ao proferido. Conclui-se que os resultados sobre a produção e a percepção indicam em conjunto a existência de um intervalo usual do ritmo ao qual os falantes aspiram ou que evitam.

Palavras-chave: ritmo, percepção da linguagem, ênfase, proeminência

Abstract: The paper presents a combined production and perception study on speech rhythm in German. The perception part shows that identifying complex rhythm patterns is only possible for speaking rates of 4-8 syll/sec. Even acoustically monotonous stimuli within this range trigger "subjective rhythms". In contrast, rhythm perception is flattened for speaking rates outside this range, irrespective of acoustic cues to rhythm. The production part accords with this finding. Speaking rates in everyday conversation vary between 4-8 syll/sec, and only fall below this range when speakers flatten their rhythm for emphatic purposes. Together, the production and perception evidence revealed a "rhythm window", which is targeted or avoided by speakers.

Keywords: rhythm; speech perception; emphasis; prominence

1. Introduction

A ‘complex rhythm’ is created by a succession of perceptually non-prominent (i.e. weak) syllables which are interrupted in more or less regular intervals by single, perceptually prominent (i.e. strong) syllables. The prominent syllables and their preceding or following non-prominent syllables together form larger perceptual units, i.e. rhythmic feet, which can, for example, be iambic, trochaic, or dactylic. Such complex rhythms fulfil many important functions in speech communication. Among others, complex speech rhythms create expectations about the upcoming prominence patterns and in this way guide listeners to the most significant pieces of information (cf. BARRY, 1981; PITT & SAMUEL, 1990; KOHLER, 2009; NIEBUHR, 2009). Moreover, complex rhythms provide a temporal framework for the production and perception of timing variables and reductions in speech, and they help listeners finding syntagmatic boundaries in utterances (cf. SMITH, CUTLER, BUTTERFIELD & NIMMO-SMITH, 1989; CUTLER & BUTTERFIELD, 1992; NAZZI DILLEY, JUSCZYK, SHATTUCK-HUFNAGEL & JUSSCZYK, 2005; DILLEY & MCAULEY, 2008; ARANTES & BARBOSA, 2010, GHITZA, 2011). Finally, rhythmic feet can facilitate and organize the processing of speech and the efficient storage of speech-transmitted information (cf. PAYNE & HOLZMAN, 1986; BROWER, 1993; MEDINA, 1994; PATEL, 1998).

However, the rhythm of speech utterances need not always be complex. Under certain conditions, utterances can also show a simple rhythmical structure. For example, in order to draw special attention to a crucial piece of information, speakers can produce the corresponding string of words in a syllable-by-syllable fashion, making each syllable similarly prominent. The result is what may be called a ‘uniform’ or ‘monotonous rhythm’, as in “*SE-VEN THOU-SAND EU-RO!*” or

"*LEAVE ME A-LONE!*". In intonation languages like English and German, this way of emphatic highlighting through a monotonous rhythm is the accent-based counterpart of lexical reduplication (cf. "*very very nice!" or "really, really big!*"). Monotonous rhythms intensify the communication channel between speakers and listeners, which is also reflected in the fact that they are typically paralleled by sustained eye contact. The underlying messages of utterances with monotonous rhythms can be paraphrased as 'I want this to be understood very clearly!' or 'mark my words!'. Rhythmic uniformity has already been implicitly described for English by HAWKINS (2003:375, cf. item 2 in Table 1) and was, more recently, addressed under the heading of persuasion by KNIGHT & CROSS (2011) AUER, COUPER-KUHLEN & MÜLLER (1999) call uniform structures like these 'Skandierungen' (scansions).

In their study on German, LANDGRAF & NIEBUHR (2011) classify rhythmic uniformity as a kind of 'emphasis for attention' (this term will also be used henceforth). They analyzed 540 utterances with this kind emphasis for attention on the basis of the KIESEL corpus ('Kieler Sammlung Expressiver Lesesprache', http://www.speechandemotion.de/ressources.htm). In the course of this analysis, LANDGRAF & NIEBUHR made a striking observation: The speaking rates of the emphasized utterances (none of which had pauses – i.e. silent intervals larger than 500 ms – between the individually accented syllables) virtually never exceeded 3.5 syllables per second (syll/sec), irrespective of the context speaking rate or the number of syllables in the utterances.

Why should there be such a speaking-rate limit? It is well known from psychoacoustic studies that sequences of acoustically identical – and hence in principle equally prominent – stimuli (e.g., clicks or tones) create so-called 'subjective rhythms' (cf. HANDEL, 1989; LARGE, 2008 for summaries). So, although the stimulus acoustics should trigger a simple, monotonous rhythm, listeners perceive a complex rhythm. The nature of such subjective rhythms is influenced by the rate with which the stimuli are presented. For example, the size of subjective rhythmic feet increases with the rate of stimulus presentation. Crucially, subjective rhythms do *not* emerge consistently when the stimuli follow one another too quickly or too slowly. Explanations for the subjective rhythm effect and its rate dependency are diverse. They are based on a change from holistic to analytic perceptual processing (cf. KOHNO, 1999) or relate to patterns/types of neural oscillation (cf. LARGE, 2008).

One wonders against this background, whether the consistently low speaking rates of the emphasized utterances in LANDGRAF & NIEBUHR (2011) were to suppress the emergence of a subjective rhythm, as this would go against the key characteristic of emphasis for attention, viz. rhythmic uniformity. This idea is not implausible, even though the stimuli used in the psychoacoustic studies on subjective rhythm differ quite a bit from speech. Psychological research (HANDEL, 1989; PATEL, IVERSEN & ROSENBERG, 2006) often emphasizes the common origin and processing of rhythm in speech and non-speech (e.g., music). For example, like musical rhythm, speech rhythm is essentially a relational phenomenon. The *relational* nature of speech rhythm is reflected in many definitions and concepts like the "*waning and waxing prominence profiles*" of KOHLER (2009:33) or the advancement in rhythm measurements from the standard deviation measures of RAMUS, NESPOR & MEHLER (1999) to the pairwise variability indices of LOW, GRABE & NOLAN (2000). It is obvious that a slower speaking rate hampers the perceptual integration and the contrastive relation of syllables and their prominence cues.

In order to further substantiate the idea that the low speaking rates of the emphatic statements were to suppress a subjective rhythm, we compared the speaking rates under emphasis for attention with speaking rates produced in a randomly chosen sub-sample of the ‘Lindenstraße’ corpus, whose 69 minutes of highly informal spontaneous German dialogues are the best representatives of natural everyday conversations within the ‘Kiel Corpus of Spontaneous Speech’ (cf. PETERS, 2001, 2005). The speaking rates were measured for the intonation phrases of the corpus. Based on the segmental and prosodic annotations, the total duration of each phrase was divided by the number of syllables within the phrase (syllables are not separately marked in the ‘Lindenstraße’ corpus, thus they had to be derived from the number of vowels / diphthongs and syllabic sonorants). We excluded interrupted phrases, phrases that contained hestitational lengthening, and phrases that consisted of less than three syllables (which were mostly used for backchanneling). The means and standard deviations for the speaking rates of the remaining 822 intonation phrases of the sub-sample are displayed in Figure 1, arranged by ascending syllable number. Figure 1 also shows the speaking-rate means and standard deviations that were found by LANDGRAF & NIEBUHR (2011) for the 540 sentences with emphasis for attention in the KIESEL corpus.

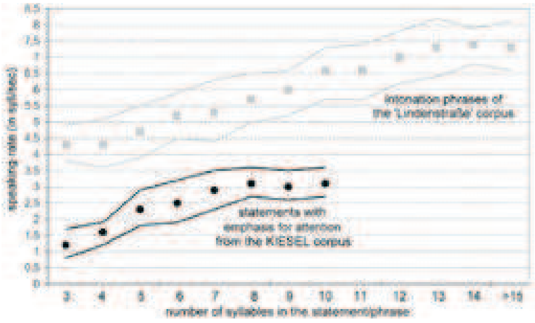

Fig. 1: Speaking-rate means (circles/squares) and standard deviations (lines) determined for 540 statements with emphasis for attention (LANDGRAF & NIEBUHR 2011, black) as well as for a sub-sample of 822 intonation phrases from the ‘Lindenstraße’ corpus (PETERS 2005).

Figure 1 reveals three important points. First, as for the 'Lindenstraße' intonation phrases, there is a clear correlation between the speaking rate of a phrase and its number of syllables (r=0.77; *df*=820; p<0.001; Pearson's coefficient). The speaking rate increases with longer phrases, which accords well with findings from other languages (cf. NAKATANI, O'CONNOR & ASTON, 1981; BRØNDSTEDT & MADSEN, 1997; QUENÉ, 2005; YUAN, LIBERMAN & CIERI, 2006). Second, within this correlation, the speaking rates vary between about 4 and 8 syll/sec. Similar variation limits are also known from other languages and corpora (cf. YUAN ET AL., 2006). However, there are no obvious reasons for these limits. For example, we can see from the statements with emphasis for attention that speech can be produced at much lower rates; and as regards the opposite end of the speaking-rate range, studies like that of DELLWO & WAGNER (2003) have demonstrated that speaking rates can be much higher than 8 syll/sec, if the speakers are instructed to speak as fast as they can. Third, there is a similar correlation of speaking rate and number of syllables for both the 'Lindenstraße' phrases and the emphasis-for-attention statements (r=0.62; *df*=538; p<0.001). However – and this is the crucial point – although the numbers of syllables per utterance *do* overlap between the two datasets, the speaking rates do clearly *not* overlap.

In view of Figure 1 and its empirical background, the present paper deals with a perception experiment that addresses the following question and assumptions:

- Do the bipartite speech-rate distributions in Figure 1 reflect a "*rhythm window*"? That is, is there a range of speaking rates in which listeners perceive complex rhythms and that is targeted in the speech of everyday conversations and deliberately avoided for particular communicative purposes like the expression of emphasis for attention? If this is the case,

- then a speaking-rate range of about 4-8 syll/sec should *facilitate* the consistent identification of acoustically triggered complex rhythms in speech stimuli as well as the emergence of subjective rhythms in acoustically uniform speech stimuli,
- whereas speaking rates below 4 syll/sec and above 8 syll/sec should *impede* the consistent identification of acoustically triggered complex rhythms in speech stimuli as well as the emergence of subjective rhythms in acoustically uniform speech stimuli.

As will be explained below, these assumptions were tested with native speakers of German by means of a 3AFC task based on reiterant speech stimuli.

2. Method

2.1 Speech material

The stimuli of the present study were developed from stimuli of a previous study by KOHLER (2008). KOHLER tested how the presence and absence of syllable-based variation in F0, duration, and intensity affected the perception of prominence patterns – and hence of rhythmic feet – in German. The stimuli in KOHLER'S experiments started from his own natural productions of the syllable <ba> ([ba̱]). He produced this syllable repetitively with trochaic or iambic rhythms, i.e. as <BAbaBAbaBAba…> or <baBAbaBAbaBA…>.

While KOHLER took only a single <BA> from the trochee productions and resynthesized all his reiterant stimuli from there, the present study used one <BAba> and one <baBA> disyllable from his original, unpublished speech material. These disyllables were selected because they represented a very basic kind of speech material: [ba̲ba̲] is a nonsense disyllable in German, which rules out lexical biases, and the trochaic or iambic feet of <BAba> and <baBA> are the two most basic and therefore probably also the most clearly perceivable forms of complex rhythms.

The selected disyllables were extracted from about the center of KOHLER'S syllable strings, in order to exclude prosodic patterns of turn boundaries. The extracted disyllables showed pronounced between-syllable differences in the three pivotal prominence cues F0, duration and intensity. Table 1 and Figures 2(a)-(b) summarize these between-syllable differences for the <BAba> and <baBA> tokens. As can be seen, the more prominent and rhythmically strong <BA> syllables were longer, louder, and spanned by high-rising F0 peaks. Apart from small parts of the rising or falling peak slopes that reached into the adjacent <ba> syllables, the latter show no separate F0 movements. Altogether, Table 1 and Figures 2(a)-(b) leave no doubt that the acoustically inherent rhythm types of the selected disyllables were firstly complex, and secondly in principle immediately identifiable as having an either trochaic (<BAba>) or iambic (<baBA>) rhythm.

	trochaic disyllable		**iambic disyllable**	
	<BA>	<ba>	<ba>	<BA>
F0 maximum (Hz)	135	100	100	130
Syllable duration (ms)	260	190	195	270
Intensity maximum (dB)	81	78	77	81.5

Tab. 1: Acoustic-prosodic characteristics of the <BAba> (left) and <baBA> (right) syllables.

While the different trochaic and iambic structures are firmly anchored in the two disyllables in terms of acoustic-prosodic cues to prominence, there were no further relevant differences between all four syllables. For example, the first three formant frequencies amounted to about 710 Hz (F1), 1,230 Hz (F2) and 2,270 Hz (F3) in the middle of all open vowels. This formant pattern is typical of male speakers in German (cf. SIMPSON, 1998). All four vowels showed a modal voice quality with a constant harmonic amplitude difference (H1-H2) of about -1 dB (cf. KLATT & KLATT, 1990 for this measure). The preceding bilabial plosives may be characterized as being widely voiced (although the vocal-fold vibrations were partly irregular). They ended in a clear release burst that was not followed by post-aspiration.

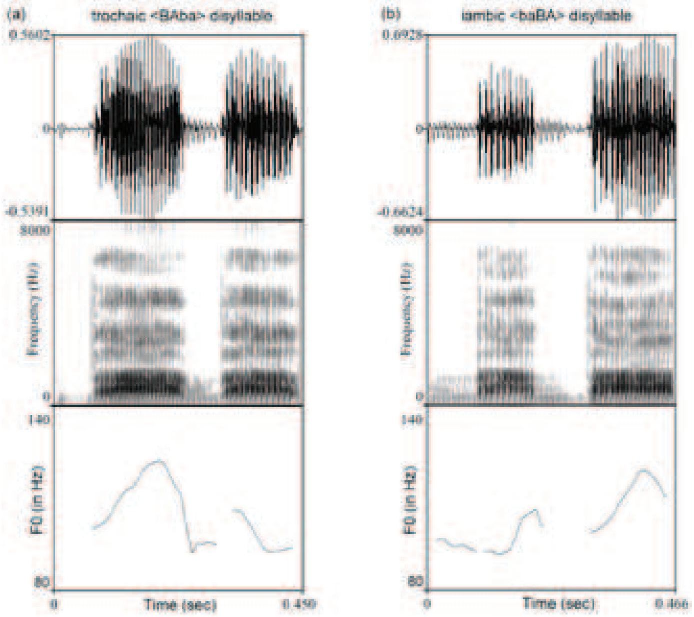

Fig. 2: Oscillogram (top), spectrogram (0-8 kHz) and F0 pattern (bottom, 80-140 Hz) of the two naturally produced disyllables <BAba> (a, trochaic) and <baBA> (b, iambic) that represent the speech material of the present study.

2.2 Stimulus generation

The first step of the stimulus generation aimed at creating a third disyllable. To this end, we took the

<BAba> disyllable, cut off the second <ba> and replaced it by reduplicating the initial <BA>. The resulting <BABA> token consisted of two phonetically absolutely identical syllables. So, in terms of acoustic cues to perceived syllable prominence and speech rhythm, the third disyllable should trigger neither a trochaic nor an iambic rhythm. Rather, <BABA> is acoustically inherently uniform and has hence the potential to create a simple, monotonous rhythm.

In the second step of the stimulus generation, each of the three disyllables (<BAba>, <baBA> and <BABA>) was copied nine times. The nine copies were then concatenated and attached to their corresponding original disyllable. This resulted in three different strings of disyllables. Each string consisted of 10 identical disyllables. We created these strings, since it was easier to perform the third step of the stimulus generation, the speaking-rate manipulation, on the basis of a larger number of syllables. Moreover, by integrating ten disyllables into a single sound file, we created a buffer that avoided disfluencies during the looped presentation of the sound files in the perception experiment.

The third and final step of the stimulus generation concerned the duration manipulations that were carried out using the PSOLA resyntheses in PRAAT (cf. BOERSMA, 2001). The manipulations were linear and oriented towards the syllables-per-second measure. With reference to the duration measurements given in Table 1, the original speaking rates of the three strings of disyllables were

- 4.44 syll/sec in the <BAba> string,
- 4.30 syll/sec in the <baBA> string,
- and 3.85 syll/sec in the <BABA> string.

The initial duration manipulation served to equalize these minor differences. To this end, the overall durations of all three disyllable strings were adjusted to 5 sec, which corresponds to a common speaking rate of 4 syll/sec. Based on this homogeneous point of departure, the overall durations of the syllable strings were subsequently increased in four and decreased in 12 steps of between 100 ms and 2000 ms. Together with the starting point of 4 syll/sec, these subsequent duration manipulations resulted in 17 speaking-rate conditions that ranged from 2 syll/sec to 10 syll/sec in equal-sized steps of 0.5 syll/sec. The two extreme conditions that were resynthesized for the <BABA> string are displayed in Figures 3(a)-(b). Since each speaking-rate condition was resynthesized as a separate stimulus, a total of 51 stimuli were created, 17 stimuli for each string of disyllables. With reference to the introduction and compared with mean values determined by DELLWO, FERRAGNE & PELLEGRINO (2006) for German, the speaking rates of stimuli 1-5 (2-4 syll/sec) may be classified as low; stimuli 6-12 represented a medium speaking-rate class (4.5-7.5 syll/sec); and the speaking rates of stimuli 13-17 (8-10 syll/sec) were high.

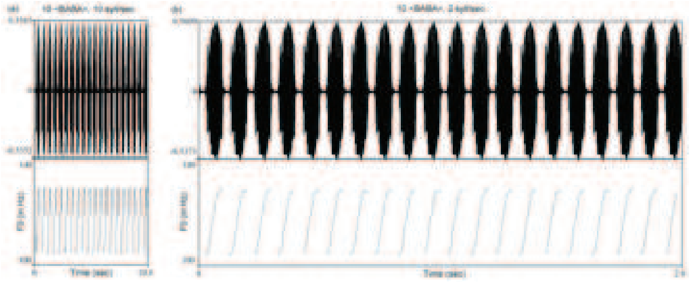

Fig. 3: Oscillogram (top) and F0 pattern (bottom) of the string of 10 <BABA> disyllables, resynthesized with speaking rates of 10 syll/sec (a, stimulus 17) and 2 syll/sec (b, stimulus 1).

2.3 Participants

A total of 16 native speakers of Standard Northern German – 10 females and 6 males – participated in the perception experiment. They were all undergraduate students of Empirical Linguistics at the University of Kiel and between 21 and 26 years old. Most of them (i.e. 11 subjects) were musically experienced in the sense that they regularly played instruments or sang in a choir/band in their spare time. However, none of the participants was an expert musician and/or had any lecture on speech prosody. All subjects reported normal hearing and received credit points for their successful participation in the experiment.

2.4 Perception experiment

The entire perception experiment consisted of 306 stimuli. This number resulted from the fact that all 51 generated stimuli (3x17) occurred six times, but in separate blocks. The stimulus order varied between the blocks as well as between the participants. The order was the result of a quasi-randomization. That is, automatically generated stimulus randomizations were selected and edited such that the frequencies with which the three stimulus types (<BAba>, <baBA>, <BABA>) and the three speaking-rate classes (low, medium, high, cf. 2.3) occurred after each other were approximately balanced across the six blocks of each participant. This balance was to compensate for possible context effects on the judgment behaviour.

The six blocks were judged by the 16 listeners in separate experimental sessions, which took place on a weekly basis (right after a lecture) in a sound-treated room of the Department of General Linguistics at the University of Kiel. Conducting the experiment blockwise with a full week time in between the individual sessions was to avoid that the subjects' responses were biased by rapidly emerging perceptual artefacts based on learning, habituation, or other kinds of desensitisation to rhythmic structures in speech.

Each session began with playing the same previously recorded oral instruction. It stated that the following experiment would be about the perception of types of speech rhythm in 51 subsequently presented segmentally constant strings of [ba_] syllables. The subjects' task would be to listen carefully to each of these strings and to specify the perceived type of speech rhythm by clicking the corresponding button on the screen in front of them (3AFC). Although each string would basically continue until a button has been clicked, judgments were to be made as quickly as possible. The whole experimental procedure (auditory stimuli, their coordination with visual response buttons, and identification of judgments) was executed with the reaction-measurement device RMG4, which has been developed and programmed at the former Kiel Institute of Phonetics and Digital Speech Processing.

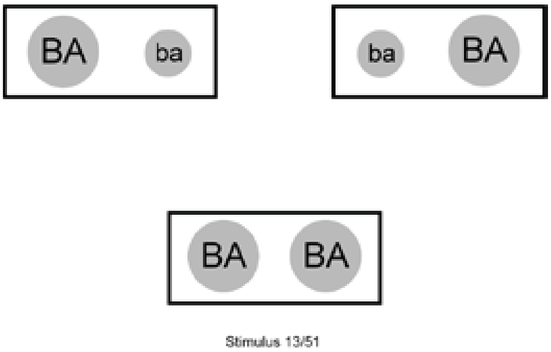

Fig. 4: Representation of the three buttons shown on the PC screen in front of the subjects during their judgment of the experimental stimuli. The button design was guided by the intonational transcriptions in the British School tradition (cf. JONES, 1959). That is, perceived prominence differences and hence rhythmic structures are reflected in the different sizes of the grey circles. The orthography (small letters vs. capital letters) supports these size differences. The three buttons have equal horizontal and vertical distances to each other.

The participants heard the instruction as well as the 51 stimuli of each session over headphones at a pre-adjusted moderate loudness level that was constant across all participants and sessions. After the instruction was over, the subjects started the experiment when they were ready by a click on a 'Start' button. Then, the screen showed three equally large buttons. They are displayed in Figure 4 in their original, evenly spaced arrangement. Placing buttons of the same shape and size in equal distances to each other on the screen was to avoid judgment biases. We further randomized across the 16 participants and experimental sessions, which of the three buttons (top left, top right, and bottom center) was associated with <BAba>, <baBA>, and <BABA> (perceived rhythm type).

The individual stimuli were introduced by a constant silent interval of 3 sec, and each stimulus stopped immediately after a button was clicked. When listening to the stimuli, the subjects did not see any stimulus information on the screen, except for their progress in terms of the current stimulus number relative to the total number of 51 (see bottom of Fig. 4).

The stimuli were played in endless loops. So after the string of ten disyllables was played, the stimulus presentation immediately started over again with the first disyllable of the string. This looping procedure did not create any perceivable discontinuity and continued until a judgment was made. This already implies that the 16 subjects performed the experimental sessions at individual paces. Yet, all participants needed a similar amount of time, i.e. between 7-9 min, to complete an experimental session.

3. Results

The results of the perception experiment are descriptively summarized in Figures 5(a)-(c). As regards inferential statistics, a series of three repeated-measures ANOVAs were conducted on the basis of the two three-level fixed factors Stimulus Inherent Rhythm (i.e. the acoustically inherent <BAba>, <baBA>, or <BABA> structures) and Speaking Rate Class (low, medium, high). Judgment frequencies per stimulus served as dependent variable. Since the 3AFC task allowed the subjects to choose between the perceived rhythm categories <BAba>, <baBA> and <BABA>, each of the three repeated-measures ANOVAs addressed a different judgment category.

Figures 5(a)-(c) illustrate in terms of percentages (of a maximum of 96 judgments), how often a particular combination of Stimulus Inherent Rhythm and Speaking Rate condition triggered the perception of trochaic <BAba>, iambic <baBA> or rhythmically uniform <BABA> disyllables. Starting with a descriptive analysis of these percentages, the first and probably most obvious result of the perception experiment concerns the perception of a uniform rhythm in the stimuli. Across the resynthesized speaking-rate continuum from 2-10 syll/sec, the <BABA> judgments take a clear <u>-like shape. So, very low speaking rates of about 2-3.5 syll/sec and very high speaking-rates of at least 8.5 syll/sec both made the subjects perceive the stimuli predominantly as <BABA>. This was true independently of the stimuli's acoustically inherent rhythm type and can hence see seen in more or less the same way in all three displays (a)-(c) of Figure 5.

Second, there was a range of speaking rates in between 3.5 and 8.5 syll/sec in which the perceived speech rhythm of the stimuli was not judged to be simply uniform. In fact, displays (a) and (b) of Figure 5 show for speaking rates from about 4 to 8 syll/sec that the perceived rhythm type accords very well (and in many cases even almost perfectly) with the acoustically inherent rhythm type of the corresponding stimulus. That is, stimuli based on the naturally produced <BAba> disyllable were also judged to have a trochaic <BAba> rhythm. Likewise, stimuli based on the naturally produced <baBA> disyllable clearly triggered iambic <baBA> perceptions. In the few exceptions (of about 10% of the cases) in which the perceived rhythm type differed from the inherent rhythm type of the stimuli, the judgments almost all refer to the respective other complex rhythm. So, the <BAba> stimuli were judged to show a <baBA> rhythm and vice versa. Cases in which <BAba> or <baBA> were presented at moderate speaking rates and still yielded rhythmically uniform <BABA> judgments were very rare. In terms of concrete numbers, there were only 79 of such cases within 1,728 judgments. This corresponds to about 4.5%.

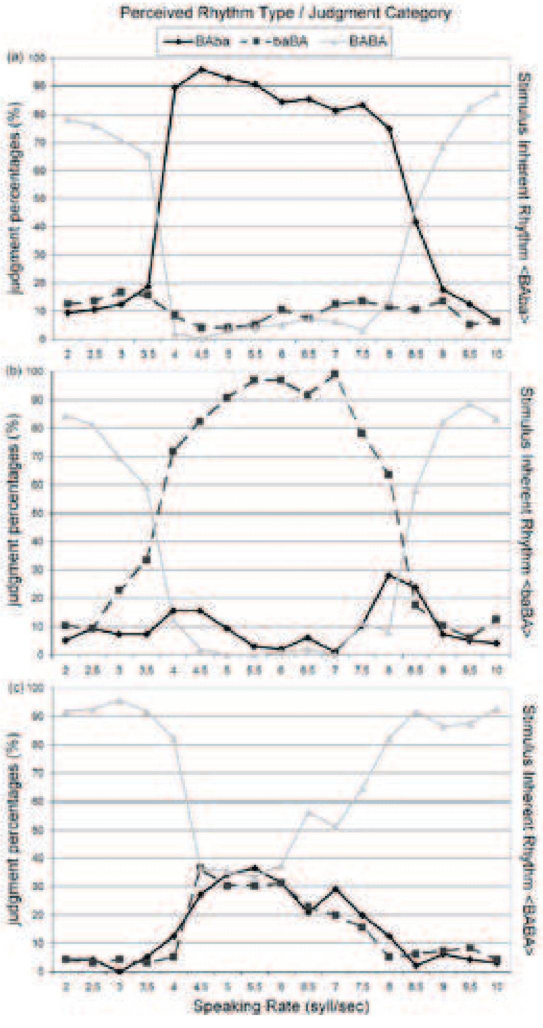

Fig. 5: Percentages of <BAba> (black line), <baBA> (dark grey line) and <BABA> (light grey line) perceptions across the 17 speaking rate conditions (x axis) that were resynthesized for the three disyllable strings based on <BAba> (a, top), <baBA> (b, middle) and <BABA> (c, bottom). Each data point represents 96 judgments.

A complex speech rhythm was not only perceived for the <BAba> and <baBA> stimuli. The third important result of the perception experiment is that also the acoustically inherently uniform <BABA> stimuli were able to trigger a non-uniform rhythm perception when being presented at moderate speaking rates. Figure 5(c) shows for speaking rates between 4.5 and 6 syll/sec that the rhythm judgments were almost equally distributed across the three judgment categories. So, only about 33% of the <BABA> stimuli were "correctly" associated with a <BABA> rhythm by the subjects. In the remaining 66% the subjects had the impression of either a trochaic <BAba> or an iambic <baBA> rhythm in the <BABA> stimuli. Up to 7.5 syll/sec, the perception of complex <BAba> or <baBA> rhythms were still quite frequent and together represented between 50% and 33% of the judgments, before the acoustically inherently uniform <BABA> stimuli also became perceptually clearly uniform again for speaking rates higher than 8 syll/sec.

In summary, the descriptive analysis of the results points to clear effects of the two factors inherent rhythm type and speaking rate on the perceived rhythm type. Table 2, which provides the key statistics of the three repeated-measures ANOVAs, confirms this assumption. The two fixed factors Stimulus Inherent Rhythm and Speaking Rate Class contributed highly significantly to explaining the variance in the data of the three judgment categories <BAba>, <baBA>, and <BABA>. However, each ANOVA also revealed a highly significant interaction between Stimulus Inherent Rhythm and Speaking Rate Class. Given the <u>-shaped judgment curves in Figures 2(a)-(c), these interactions were not surprising. They were taken into account by splitting the two factors and running separate repeated-measures ANOVAs for each factor. It was considered sufficient in this context to perform only one (rather than three) repeated-measures

ANOVA for each factor. The one-way ANOVA for the Stimulus Inherent Rhythm was based on the <BAba> judgments. The one-way ANOVA for the Speaking Rate Class used the <BABA> judgments.

	Perceived Rhythm Type / Judgment Category		
	<BAba>	<baBA>	<BABA>
Stimulus Inherent Rhythm (<BAba>, <baBA>, <BABA>)	F(2,30) = 1,256.80 $p<0.001$; $\eta^2_p = 0.99$	F(2,30) = 1,145.19 $p<0.001$; $\eta^2_p = 0.99$	F(2,30) = 274.99 $p<0.001$; $\eta^2_p = 0.94$
Speaking Rate Class (low, medium, high)	F(2,30) = 544.64 $p<0.001$; $\eta^2_p = 0.97$	F(2,30) = 497.77 $p<0.001$; $\eta^2_p = 0.97$	F(2,30) = 1,786.47 $p<0.001$; $\eta^2_p = 0.99$
interaction Stimulus Inherent Rhythm x Speaking Rate Class	F(4,60) = 304.01 $p<0.001$; $\eta^2_p = 0.95$	F(4,60) = 252.98 $p<0.001$; $\eta^2_p = 0.94$	F(4,60) = 136.35 $p<0.001$; $\eta^2_p = 0.90$

Tab. 2: Results summary of the three repeated-measures ANOVAs that were performed for each type perceived rhythm (columns) on the basis of two three-level factors Stimulus Inherent Rhythm and Speaking Rate Class (rows). Values for F (*dfs* in brackets), alpha-error probabilities (p), and partial eta-squared (η^2_p, i.e. estimated effect sizes) are shown.

The one-way ANOVA for the Stimulus Inherent Rhythm again yielded a highly significant main effect of this factor (on the <BAba> judgments) [F(2,30)=1,256.80; $p<0.001$; $\eta^2_p = 0.98$]. Multiple comparisons (with Bonferroni correction) showed additionally that this overall significance stems from significant differences (of $p<0.01$) between all three factor levels – <BAba>, <baBA>, and <BABA>. This is due to the fact that the <BAba> stimuli yielded a lot more <BAba> judgements than the <BABA> stimuli, which in turn yielded more <BAba> judgments than the <baBA> stimuli (cf. Fig 5a-c). A similarly clear picture emerged for the other ANOVA that was concerned with Speaking Rate Class. That is, the low, medium and high speaking-rate classes had a highly significant main effect on the <BABA> judgments [F(2,30)=2,493.64; $p<0.001$; $\eta^2_p = 0.99$]. This effect rested on significant differences (of $p<0.001$ after Bonferroni correction) between all three speaking-rate classes. These differences are clearly reflected in the <u>-shaped judgment curves in Figures 5(a)-(c).

4. Conclusion and Outlook

The results of the perception experiment provided evidence that there was a particular range of speaking rates

- (a) for which the acoustically triggered, inherent rhythm type of the stimuli was identified by listeners with a high degree of consistency, and
- (b) for which even the acoustically inherently uniform <BABA> stimuli triggered the perception of a “subjective rhythm”, in the form of non-uniform trochaic or iambic <BAba> or <baBA> patterns.

Stimuli with speaking rates outside this particular range were predominantly perceived to be rhythmically uniform <BABA> sequences. This was even true for those stimuli whose acoustic-prosodic patterns clearly cued a complex speech rhythm (it should be noted for the sake of completeness that some subjects reported to hear the [ba_] syllables as [da_] when they were presented at very high speaking rates; however, these subjects assured at the same time that this rate-dependent sound change had not influenced their rhythm judgments).

It may be possible that the overall results pattern is to some extent influenced the 3AFC task, which ruled out responding with any other complex rhythmic grouping than trochee and iamb. In particular, these two categories may have been insufficient for the large spectrum of possible “subjective rhythms” that may have emerged, although even on request no subject noted a lack of judgment categories. Irrespective of this potential shortcoming, there can be no doubt that the 3AFC task was able to capture the crucial perceptual difference, namely that between complex and uniform rhythms.

So, in summary, the experiment lends strong support to the existence of a "*rhythm window*" in speech communication. Those speech utterances that are supposed to be rhythmically complex – which is advantageous for a number of communicative and cognitive reasons and hence the default case in speech production (cf. introduction) – must fall within this window. More specifically, this means that their speaking rates should be between about 4-8 syll/sec. The significance of this claim is underlined by the analyzed subset of intonation phrases from the spontaneous 'Lindenstraße' dialogues, whose naturally produced speaking rates agree very well with the perceptually determined "rhythm window" (cf. Fig.1). There were only a few phrases in the 'Lindenstraße' subset with speaking rates below 3.5 and above 8.5 syll/sec. On average the speaking rate was always higher than 4 syll/sec and lower than 7.5 syll/sec. The opposite applies to the statements from the KIESEL corpus, which were supposed to be rhythmically uniform in order to express emphasis for attention. Their speaking rates did never exceed 4 syll/sec, independent of utterance length. The average speaking rate even stagnated already at about 3 syll/sec (cf. Fig.1). The vast majority of stimuli presented at about this rate in the perception experiment were judged to be rhythmically uniform. So, from a perceptual point of view speaking rates below 4 syll/sec are sensibly chosen for conveying emphasis for attention.

The combined production and perception evidence suggests that speakers are aware of the "rhythm window", and that the speaking rates of utterances in everyday conversation are deliberately targeted to fall within or outside the "rhythm window" in order to enhance their rhythmic complexity or uniformity. This suggestion implicitly assumes a correspondence between the acoustic cues to prominence or rhythm on the one hand and the speaking rate on the other. That is, utterances that fall within/outside the "rhythm window" are also acoustically designed to trigger complex/uniform prominence and rhythm patterns. Scrutinizing this assumption could be a task of follow-up studies.

Moreover, the present paper introduced, analyzed and discussed the "rhythm window" from a one-dimensional perspective, with speaking rate as the only determining parameter. However, it seems well possible that the "rhythm window" is actually a multidimensional phenomenon in the sense that variations along other acoustic or perceptual parameters also facilitate or impede the creation and identification of speech rhythm, in this way interacting with speaking rate. This possibility could as well be an interesting subject of follow-up studies. Such studies should include poetry readings, since poems can be performed at very low speaking rates and still create strong rhythmic impressions (cf. WAGNER, 2012). Maybe poetry rhythms at very low speaking rates use particular compensation strategies like timing changes that maintain or enhance the relational, syntagmatic contrasts on which prominence and grouping perceptions rely, and that are not at work in everyday communication (because they are inefficient or unnecessary). It fits in with this assumption that, impressionistically, the speaking rate in rhythmically performed poems can only be reduced at the expense of fluency – another perceptual quality of rhythmic speech whose understanding is still in its infancy (cf. NIEBUHR & WOLF, 2011; DILLEY, WALLACE & HEFFNER, 2012).

Finally, even though the cross-linguistically limited speaking-rate variation of naturally produced utterances and the general cognitive mechanisms of prominence and rhythm perception suggest that every language has a "rhythm window", it must still be regarded as an open question, whether every language has the same "rhythm window" in terms of both determining parameters and their crucial ranges. For example, since each language has specific speaking-rate and syllable-structure patterns and more or less restricts the exploitation of prominence cues for rhythmic purposes, it is likely that the "rhythm window" – just as any other aspect of the speech code – is to a certain degree language specific. In any event, the notion of a "rhythm window" has added another important facet to our advancing but still insufficient efforts of measuring and conceptualizing speech rhythm and its relations to various aspects of the timing and cognitive processing of speech.

Acknowledgements

The author would like to thank Klaus Kohler for the many inspiring discussions on the nature of speech rhythm and for sharing his speech stimuli with me. Further thanks are due to Elmar Eggert and Monique Fritscher for preparing the Portuguese counterpart of my English abstract. Last but not least, I am greatly indebted to my two anonymous reviewers for their helpful comments and suggestions.

References

ARANTES, P.; BARBOSA, P. Production–perception entrainment in speech rhythm. *Proceedings of the 5th International Conference of Speech Prosody, Chicago, USA*, p. 1-4, 2010

AUER, P; COUPER-KUHLEN, E.; MÜLLER, F. *Language in Time. The Rhythm and Tempo of Spoken Interaction*, Oxford: Oxford University Press, 1999.

BARRY, W.J. Prosodic functions revisited again! *Phonetica*, 38, p. 120-134, 1981.

BOERSMA, P. Praat, a system for doing phonetics by computer. *Glot International*, 5, p. 341-345, 2001.

BRONSTED, T.; MADSEN, J.P. Analysis of speaking rate variation in stress-timed languages. *Proceedings of the 1st Eurospeech Conference, Rhodes, Greece*, p. 481-484, 1997.

BROWER, C. Memory and the Perception of Rhythm. *Music Theory Spectrum*, 15, p. 19-35, 1993

CUTLER, A.; BUTTERFIELD, S. Rhythmic cues to speech segmentation – Evidence from juncture misperception. *Journal of Memory & Language*, 31, p. 218-236, 1992.

DELLWO, V.; WAGNER, P. Relationships between speech rate and rhythm. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain*, p. 471-474, 2003.

DELLWO, V.; FERRAGNE, E.; PELLEGRINO, F. The perception of intended speech rate in English, French, and German by French speakers. *Proceedings of the 3rd International Conference of Speech Prosody, Dresden, Germany*, p. 217-220, 2006.

DILLEY, L.C.; MCAULEY, J.D. Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, 59, p. 294–311, 2008.

DILLEY, L.C.; WALLACE, J.; HEFFNER, C. Perceptual isochrony and fluency in speech by normal talkers under various task domains. In: NIEBUHR, O. (Org.). *Understanding Prosody – The role of context, function and communication*. Berlin/New York: de Gruyter, 2012, p. 237-258.

GHITZA, O. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, 2, p. 1-13, 2011.

HANDEL, S. *Listening – An introduction to the perception of auditory events*. Cambridge: MIT Press, 1989.

HAWKINS, S. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31, p. 373-405, 2003

JONES, D. *An Outline of English Phonetics*. Leipzig: Teubner, 1959.

KLATT, D.H.; KLATT, L.C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87, p. 820-857, 1990.

KNIGHT, S.; CROSS, I. Rhythms of persuasion: The perception of periodicity in oratory. *Proceedings of the Perspectives on Rhythm and Timing Workshop (PoRT), Glasglow, Scotland*, p.27, 2011.

KOHLER, K.J. The perception of prominence patterns. *Phonetica*, 65, p. 257-269, 2008.

KOHLER, K.J. Rhythm in speech and language - A new research paradigm. *Phonetica*, 66, p. 29-45, 2009.

KOHNO, M. Rhythmic patterns in languages and psychology of speech perception. *Psycholinguistics on the threshold of the year 2000 : Proceedings of the 5th International Congress of the International Society of Applied Psycholinguistics, Porto, Portugal*, p. 793-797, 1999.

LANDGRAF, R.; NIEBUHR, O. *Emphasis for attention in German – An initial phonetic analysis*. Unpublished manuscript, University of Kiel, 2011.

LARGE, E.W. Resonating to musical rhythm: Theory and experiment. In: GRONDIN, S. (Org.). *Psychology of time*. Bingley: Emerald, 2008, p. 189-232.

LOW, E.L.; GRABE, E.; NOLAN, F. Quantitative characterizations of speech rhythm: Syllable timing in Singapore English. *Language and Speech*, 43, p. 377-401, 2000.

MEDINA, S. The impact of rhythm upon verbal memory. *Forefrontpublishers, 1994. Available at: http://www.scribd.com/doc/48547756/The-Impact-of-Rhythm-Upon-Verbal-Memory. Acessed in: 11 Nov. 2012.*

NAKATANI, L.H.; O'CONNOR, J.D.; ASTON, C.H. Prosodic aspects of American English speech rhythm. *Phonetica*, 38, p. 84-106, 1981.

NAZZI, T.; DILLEY, L.C.; JUSCZYK, A.M.; SHATTUCK-HUFNAGEL, S.; JUSCZYK, P.W. English-learning Infants' Segmentation of Verbs from Fluent Speech. *Language and Speech*, 48, p. 279-298, 2005.

NIEBUHR, O. F0-based rhythm effects on the perception of local syllable prominence. *Phonetica*, 66, p. 95-112, 2009.

NIEBUHR, O.; WOLF, A. Low and High, Short and Long by Crook or by Hook? *Proceedings of the 12th Interspeech Conference, Florence, Italy*, p. 1869-1872, 2011.

PATEL, A.D. Processing prosodic and musical patterns: a neuropsychological investigation. *Brain and Language*, 61, p. 123-144, 1998.

PATEL, A.D.; IVERSEN, J.R.; ROSENBERG, J.C. Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of America*, 119, p. 3034-3047, 2006.

PAYNE, Jr. M.C.; HOLZMAN, T.G. Rhythm as a factor in memory. In: EVANS, J.; CLYNES, M. (Orgs.). *Rhythm in Psychological, Linguistic and Musical Processes*. Springfield: Charles C. Thomas, 1986, p. 41-54.

PETERS, B. 'VideoTask' or 'Daily Soap Scenario' - A new method for the controlled elicitation of spontaneous speech. *IPdS, University of Kiel, 2001. Available at: http://www.ipds.uni-kiel.de/pub_exx/bp2001_1/ Linda21.html. Acessed in: 3 Jan. 2012.*

PETERS, B. The Database 'The Kiel Corpus of Spontaneous Speech'. *AIPUK*, 35a, p. 1-6, 2005.

PITT, M.A.; SAMUEL, A.G. The use of rhythm in attending to speech. *Journal of Experimental Psychology: Human Perception and Performance*, 16, p. 564-573, 1990.

QUENÉ, H. Modeling of between-speaker and within speaker variation in spontaneous speech tempo. *Proceedings of the Interspeech Conference, Lisbon, Portugal*, p. 2457-2460, 2005.

RAMUS, F.; NESPOR, M.; MEHLER, J. Correlates of linguistic rhythm in the speech signal. *Cognition*, 72, p. 1-28, 1999.

SIMPSON, A.P. Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung. *AIPUK*, 33, p. 1-233, 1998.

SMITH, M.R.; CUTLER, A.; BUTTERFIELD, S.; NIMMO-SMITH, I. The perception of rhythm and word boundaries in noise-marked speech. *Journal of Speech and Hearing Research*, 32, p. 912-920, 1989.

WAGNER, P. Meter specific timing and prominence in German poetry and prose. In: NIEBUHR, O. (Org.). *Understanding Prosody – The role of context, function and communication*. Berlin/New York: de Gruyter, 2012, p. 219-236.

YUAN, J.; LIBERMAN, M.; CIERI, C. Towards an Integrated Understanding of Speaking Rate in Conversation. Proceedings of the Interspeech Conference, Pittsburgh, USA, p. 541-544, 2006.

[Recebido em 31 de julho de 2013
e aceito para publicação em 14 de novembro de 2013]