Note: To appear in *Musicae Scientiae* (2009): Special Issue *Music and Evolution* (ed. O. Vitouch & O. Ladinig). In case of any discrepancy with the printed version, the printed version will be the 'authorized' version.)

Some parallels between language and music from a cognitive and evolutionary perspective

Gertraud Fenk-Oczlon
Department of Linguistics and Computational Linguistics
University of Klagenfurt
9020 Klagenfurt, Universitätsstrasse 65-67
gertraud.fenk@uni-klu.ac.at

August Fenk
Department of Media and Communication Studies
University of Klagenfurt
9020 Klagenfurt, Universitätsstrasse 65-67
august.fenk@uni-klu.ac.at

Abstract

Parallels between language and music are considered as a useful basis for examining possible evolutionary pathways of these achievements. Such parallels become apparent if we compare clauses and syllables in language with phrases and notes in music: Clauses as well as musical phrases typically span about 2 sec and about 5 to 10 pulses, *i.e.,* syllables or notes. The n of syllables per clause or intonation unit also can be used as a measure of tempo across languages and thus also as a means for a better understanding of typological co-variations in the rhythm of speech and music. Further correspondences were found between the size of the sound-relevant inventories, *i.e.,* vowels and musical intervals: a minimum of roughly 3 and a maximum of roughly 12 elements as well as a frequency peak at 5 elements. A link between vowels and musical intervals is also indicated by our findings that in Alpine yodellers the vowels are highly correlated to melodic direction according to their F2 ordering.
These parallels are discussed from an evolutionary perspective that either sees music as a precursor of language or both language and music as descendents of a common, "half-musical" precursor (Jespersen, 1895; Brown, 2000). A rather simple explanation of the parallels is reported: If *singing* in a broader sense of the word is the most original form of music, then the functionality of any mechanism involved in the programming and the online-control of intonation units will be reflected in language as well as in music.

1 Introduction

The evolution of language, the evolution of music, and possible connections between the evolutionary pathways of these achievements are hotly debated topics in contemporary

anthropology. We think that a thorough study of correspondences between recent language systems and recent musical systems is a useful or even necessary basis for an examination of possible reconstructions of such evolutionary pathways.

Let us begin with some well known but rather general parallels: Both language and music are organized temporally, both show rhythm and intonation and have syntactically structured sequences. In both we perceive the sounds as a sequence of pulses - in language as syllables, in music as notes. And music is, like language, "generative" in the sense that it uses rule-governed combinations of a limited number of elements to generate an unlimited number of hierarchically structured signals (Fitch, 2006, p.178, referring to Merker, 2002).

Our search for a more detailed picture of such correspondences starts (in Section 2) with language universals - *language universals* in the sense of Greenberg (1968), *i.e.,* in the sense of statistical, cross-linguistic regularities – concerning the rhythmic organization of linguistic utterances. This main part includes statistical findings by the authors (Section 2.1 and 2.2) as well as some considerations on the underlying cognitive mechanisms (2.4). In a second step we try to relate rhythm patterns in speech to relevant patterns reported by musicologists (Section 3). The comparison following in Section 4 between vowels in language and musical intervals offers a solution of the problem (*e.g.,* Rakowski, 1999) how to relate the segmental inventories of the two systems.

In both the comparisons between language and music regarding their rhythm organization and their inventories, the syllable plays a central role: It is, first of all, the rhythmically most relevant subunit of phrases, in speech as well as in singing, and irrespective of whether one sings meaningful texts or meaningless syllables. Therefore the syllable is, unlike the word, also an appropriate unit for cross-linguistic comparisons of rhythm (*cf.,* Section 2.2). And the core of (almost) any syllable is the vowel that carries most of the sound or sonority in both speech and vocal music. Last but not least the syllable seems to play a key-role in language variation (*e.g.,* Fenk-Oczlon & Fenk, 2005). Consequently, it also plays a prominent role in our considerations regarding the evolutionary pathways of language and music and the perceptual/cognitive processes involved in the perception and production of linguistic and musical utterances.

2 Rhythm and Tempo in Language

In his article on "Rhythm and Tempo", Fraisse (1982) offers two definitions of *tempo*:

> One of the perceptual aspects of rhythmic organization is *tempo*. It can be lively or slow. It corresponds to the number of perceived elements per unit time, or to the absolute duration of the different values of the durations. Evidently one passed from a definition based on frequency to a definition based on duration. […] The possibility of rhythmic perception depends on tempo, because the organization of succession into perceptible patterns is largely determined by the law of proximity. When the tempo slows down too much, the rhythm and also the melody disappear. (Fraisse, 1982, p. 151)

Like Fraisse, we will use both definitions of *tempo:* the frequency-based definition in Section 2.1, and comparisons with durational measures in Section 2.2 and 2.3. But what are the "perceived elements" of speech? The phonemes, the syllables, the words, or even bigger units than the word? And what should be considered the superordinate pattern uniting such elements? The tempo is the relevant aspect at least within temporal units or patterns separated by larger pauses. If there is no such pause, subjective grouping appears. When the temporal patterns

> are quite long, they often split up into several subunits. A pattern of six sound taps is often decomposed into two subunits of 3 + 3, of 4 + 2, or of 2 + 2 + 2 as the case may be. […] This type of analysis explains, we think, certain groupings that intervene when models have eight or ten sounds, as in research such as Garner's. (Fraisse, 1982, p. 168)

The "mono-clausal sentence", and the clause in general, is not only a language universal but can be identified as the superordinate unit – the "intonation unit" (Chafe, 1987) - followed by a pause. And the tempo within this unit is mainly conveyed by the number and the size of the syllables within the clause. The word, however, is ineligible as a "perceivable element" of a rhythmic pattern - at least in crosslinguistic comparison and because of its immense variability in duration and complexity, ranging from the monosyllabic word, which seems to be very frequent especially in isolating/analytic languages, to polysyllabic words with high syllable numbers (*e.g.,* in compounds) and to a whole sentence in some polysynthetic languages.

2.1 The size of clauses in terms of syllables

In our previous studies, the temporal unit is the clause and its "perceivable element" is, first of all, the syllable. Determining the size of a unit in terms of the number of its elements obviously corresponds to Fraisse's frequency based definition of tempo. Some results of this series of cross-linguistic studies:

a) The first of these studies (Fenk-Oczlon, 1983) started from the hypothesis that in any natural language the mean number of syllables per simple declarative sentence would be located within Miller's (1956) range of 7 plus minus 2. Simple declarative sentences encoding one proposition within one intonation unit seem to be a language universal. Native speakers of 27 typologically different languages were asked to translate a set of 22 German "mono-clausal" sentences of this sort into their mother tongue and to determine the number of syllables of each of the sentences produced. The written translations allowed the enumeration of the number of words per clause, and the number of phonemes was determined with the help of the subjects and of grammars of the respective languages. The result was, as expected, a mean number of about 7 syllables per sentence, ranging from 5 in Dutch to 10 in Japanese. Incidentally: The mean number of words in these sentences was about 4, ranging from 2.5 in Arabic to 5.4 in Chinese. This corresponds to Cowan's (2001) magical number of 4 plus minus 1 in short term memory.

b) Time-related constraints are indicated by significant cross-linguistic correlations. *E.g.,*: The more syllables per sentence, the fewer phonemes per syllable (Fenk-Oczlon & Fenk, 1985). Dutch and Japanese marked not only the endpoints of the dimension "n of syllables per sentence" (see a), but also of the dimension "n of phonemes per syllable": Dutch showed the highest, Japanese the lowest syllable complexity.[1] The negative correlation between syllable complexity and the number of syllables per sentence was, as far as we can see, the first "cross-linguistic correlation" in the literal sense of the word, *i.e.,* in the sense of a computation where each language is represented by a single data pair (*e.g.,* x = mean n of syllables per sentence, y = mean n of phonemes per syllable). A whole set of significant and mutually dependent correlations of this sort followed in a later study (Fenk & Fenk-Oczlon, 1993) and was confirmed in a somewhat extended sample of 18 Indo-European and 16 non-Indo-European languages from all continents except Australia (Fenk-Oczlon & Fenk, 1999):

The more syllables per clause, the fewer phonemes per syllable.
The more words per clause, the fewer syllables per word.
The more syllables per clause, the more syllables per word.
The more syllables per word, the fewer phonemes per syllable.

Such correlations suggest complexity trade-offs within the language system, providing a rather constant size of clauses irrespective of the rhythm type of language.

2.2 Typological differences in tempo and rhythm: Different measures and new results

In this section we do not present a direct linguistic counterpart of musical entities or regularities. Instead, we report some new statistical findings which suggest a re-interpretation of the well-known classification of languages as stress-timed, or syllable-timed, or mora-timed. The indications for or findings of typological co-variations between language and music (Section 3.2.) usually refer to these rhythm classes.

It seems to be widely accepted that languages can be classified into two (or three) rhythm classes: Stress-timed (*e.g.,* Dutch, English), syllable-timed (*e.g.,* Italian, Spanish), or mora-timed (Japanese). But it turned out to be difficult to set out clear rules for assigning a language to such categories (*e.g.,* Cummins, 2002). French is only one of many problematic cases (*cf.,* Wenk & Wioland, 1982). The isochrony hypothesis (Pike, 1945) attributes equal intervals between prominent syllables to stress-timed languages and equal syllable duration to syllable-timed languages. This distinction has never been confirmed, and the respective classification in its original form seems to be obsolete. What is not obsolete is the search for more appropriate operationalizations and criteria distinguishing between different (classes or types of) languages with respect to their rhythm organization. Recent work (*e.g.,* Ramus, 2002; Grabe & Low, 2002) mainly focuses on durational patterns of vocalic and intervocalic intervals and their variability. But measurements of the duration of *e.g.,* syllables are problematic. In order to get comparable and characteristic values for single languages one needs for instance "a variety of speakers for each language" and a "control for speech rate" (Ramus, 2002). In his paper, Ramus suspects that differences in rhythm might be closely related to differences in speech rate but that it is "almost illusory" to find a valid measure of speech rate across languages.

But actually, our previous studies mentioned in 2.1 depended on such a measure. The basic idea was the use of a **controlled set of propositions** and the comparison of the different translations with respect to their complexity at different levels: n of phonemes per syllable, n of syllables per word and per clause, n of words per clause. The number and the size of syllables are the relevant parameter that allows – without any measurements of duration! – a comparison of rhythm and tempo across languages: The smaller the syllables and the higher their number per clause, the higher the tempo in the respective language. Typical stress-timed

languages show a low tempo because of a low number of syllables or pulses per clause (Fenk-Oczlon & Fenk, 2006). Their high mean syllable complexity also allows a high variability of syllable size. Typical syllable-timed languages, however, exhibit high tempo because of a high number of rather simple syllables. Japanese occupies an extreme position within this "high frequency band" of syllable-timed languages. (In view of this extreme position of Japanese several authors suggest a separate "mora-timed" rhythm class.) The high tempo in Japanese, in connection with the restricted variability of the syllable size, makes it sound staccato-like in the ears of speakers of stress-timed languages.

Grabe and Low (2002) measured the duration of vowels and of the intervals between vowels in order to determine the "durational variability" of speech in different languages. Despite the fact that they recorded only one speaker from each language, a comparison of our complexity measures with the results of their "normalised Pairwise Variability Index (nPVI)" shows remarkable coincidences. Table 1 comprises the values of what Grabe and Low call "prototypical" languages of the three rhythm classes and, in addition, of Thai.[2] Thai is rather stress-timed (Grabe & Low, 2002, referring to Luangthongkum, 1977) and turned out to be the high-scorer in the nPVI, even above Dutch, German, and British English.

Table 1: Values for stress-timed languages (1 – 4), syllable-timed languages (5 – 6), and mora-timed Japanese (7), ordered along the continuum of syllable complexity (column b). Durational values (column c) from Grabe & Low (2002).

	a syll/clause	**b** phon/syll	**c** vocalic nPVI
(1) Dutch	5.05	2.97	65.5
(2) German	5.50	2.84	59.7
(3) English	5.77	2.69	57.2
(4) Thai	5.29	2.51	65.8
(5) French	5.32	2.47	43.5
(6) Spanish	7.96	2.09	29.7
(7) Japanese	10.23	1.88	40.9

Both our complexity measures and the durational variability measure by Grabe and Low show, instead of a borderline between separate rhythm classes, a continuum between the prototypical stress-timed languages and the prototypical syllable-timed languages and (or inclusive of) mora-timed rhythm.

We assumed correlations between all three measures (three columns in Table 1). But we have to add that a correlation between a and b corresponds to our "old" (Fenk-Oczlon & Fenk, 1985) correlation mentioned in Section 2.1; the only thing surprising in this respect is the fact that this coefficient proved to be highly significant in a sample of only 7 languages. We expected, moreover, that syllable complexity (column b) as the apparently most relevant parameter in language variation would explain most of the variance. The results conform to expectations: The correlations of b with a ($r = -.89$, $p < .01$) and with c ($r = +.84$, $p < .05$) were higher than the correlation between a and c ($r = -.69$, not significant). And while the partial correlations r_{ab-c} and r_{bc-a} revealed relatively high coefficients (.77 and .58 respectively), the partial correlation excluding the syllable complexity (r_{ac-b}) was near zero (.08).[3]

The classification problems mentioned above as well as our correlational findings suggest a re-interpretation of the respective "rhythm classes": They are not distinctive categorical classes, but at best types that can be assigned more or less loosely to the tempo-related continua in our Table 1. In cases of conflict we would rely, after all, on variable b, *i.e.,* syllable complexity.

In any case it might be an interesting question for future research to determine whether there is a cross-cultural correlation between the number of syllables per clause and, at least in folksongs, the number of notes per phrase. Of those languages listed in Table 1, one should expect Japanese to show the highest number of notes per phrase and per unit of time.

2.3 The duration of intonation units and syllables

Studies measuring the duration of intonation units are rather rare: According to Chafe (1987, p. 22), "new intonation units typically begin about two seconds apart. Evidently active information is replaced by other, partially different information at approximately two second intervals". In Finnish Määttä (1993) found a mean length of breath groups in the region of 2.1 to 2.2 sec, and of 3.2 to 3.3 sec inclusive pauses.

Martin (1972) assumes that a rhythmic pattern or a prosodic unit consisting of up to seven syllables often corresponds to breath groups. But what is the duration of these prosodic units

and of their basic element, the single syllable? If one takes a mean 2 sec-duration of the clauses for granted, the duration of the syllable can be estimated simply mathematically with the number of syllables obtained in our cross-linguistic studies. This results in a mean syllable duration of 200 msec in Japanese up to 400 msec in Dutch.

The numbers of syllables taken as the basis of this estimation (10 in Japanese, 5 in Dutch) are obtained from simple but complete declarative sentences such as *the sun is shining* or *the spring is on the right*. In common oral discourse one might however suspect a different situation, *e.g.,* dialogs interspersed with "shorter" intonation units, sometimes comprising only 1 or 2 syllables, and less complex syllables in casual as compared with formal language. Gigler (in preparation) studied 9 dialogues (Carinthian dialect; a total of 1055 intonation units) and found a mean length of 6.04 syllables and 1.373 seconds per intonation unit, *i.e.,* a mean of 227 msec per syllable.

Chu and Feng (2001) analyzed a huge corpus containing 13,000 sentences of Mandarin speech. They report 245 msec as the overall mean duration of syllables; more than 99.5 % of all syllables were between 100 and 500 msec. And Kegel (1990) concludes from a series of psycholinguistic experiments that 100 to 500 msec are necessary for the processing of syllables.

2.4 Cognitive constraints forcing language universals

Our working memory seems to be limited in terms of the number of units that a subject can handle - *cf.,* Miller's (1956) "magical number 7 plus or minus 2" or Cowan's (2001) "magical number 4" – as well as in terms of duration – *cf.,* a "psychological present" of roughly 2 sec in Fraisse (1982) or a similar span in Baddeley's (1986) phonological loop model. In Fraisse's own words:

> In order to understand this, let us take the example of the tick-tock of a clock. The sounds are linked together in groups of two. Let us suppose that one can slow down this tick-tock indefinitely. There comes a moment when the tick and the tock are no longer linked perceptually. They appear as independent events. This upper limit is also that where all melody disappears, and is substituted by isolated notes. The limit proposed by Bolton (1580 msec) is without doubt too precise. MacDougall rightly situated it between 1500 and 2000 msec. We propose retaining a value of about 1800 msec. Beyond this duration subjective rhythmization becomes impossible. (Fraisse, 1982, p. 156)

A time span of 2 sec seems to be critical in psychophysics as well. Lavoie and Grondin (2004, p. 198) found that, contrary to the constant predicted by Weber's law, the Weber fraction is larger at 2 sec than at 0.2 sec and argue that this might "be due to the fact that 2 s is beyond a temporal span limit for processing information."

Some authors such as Wittmann and Pöppel (1999-2000) tend to localize the size of the relevant span in the region of 2-3 sec. In this context, Schleidt and Kien (1997, p. 98) argue that the turn over time in the Necker Cube effect, which is already mentioned in Fraisse (1985, p. 96) as indicating the size of the psychological present, lies, according to more recent studies, "within a few seconds with a peak around 3 seconds".

If such constraints force the language universals reported above, then they should be responsible for corresponding universals in music, too.

3 Parallels in music?

3.1 Syllables and clauses in language – notes and phrases in music

Some parallels between language and music become apparent if we compare syllables with notes and clauses with musical phrases:

a) The duration of musical phrases roughly corresponds to Fraisse's psychological present (Parncutt & Pascall, 2002).

b) The number of musical pulses is located within a range of 30 to 300 beats per minute (Parncutt & Drake, 2001, referring to Fraisse, 1982).

This would again amount to a maximum of 10 pulses within a span of 2 seconds (300 pulses per min = 5 pulses per sec = 200 msec per pulse).

c) The performance of musicians seems to reflect working memory limits.

Sloboda (1982, p. 485) reports that the eye-hand span (EHS), *i.e.*, the amount of reading ahead, "of good readers was typically six or seven notes while that of poorer readers was only three or four notes." And typically, "a good sight-reader will execute a note about 2 sec after reading it." The EHS "decreased for meaningless (atonal) material and showed a tendency to expand or contract with phrase boundaries."

d) 8 or 9 pulses per musical phrase is a widespread pattern.

In Huron's (1996) analysis of the Essen Folksong Collection, a corpus of more than 6000 mostly European folksongs, "all phrases were extracted from the database and sorted according to the number of untied notes in the phrase. Tied-notes were treated as single notes and rests were ignored." Using this method, he found 8 notes as the most common and as the

median phrase length. Half of the phrases were 7 – 9, and three-quarters 6 – 10 notes in length.

Temperley (2001, p. 69) analyzed the Ottmann collection and found a mean of 7.5 notes per phrase; "over 75 % of phrases have from 6-10 notes, and less than 1 % have fewer than 4 or more than 14." In the Essen Folksong Collection, he found a slightly larger mean of 9.2 notes per phrase. And in polyphonic music, such as the Mozart String Quartet K. 387, he found not only phrases with far more than 8 notes but also cases where it was problematic to determine phrase length in terms of number of notes. "One possible solution would be to express the 'optimal length' of a phrase in terms of absolute time, rather than number of notes; in those phrases containing many notes, the notes are usually very fast." (Temperley, 2001, pp. 82-83)

3.2 Typological covariations between language and music?

Are there any results pointing to typological co-variations between language and music? If so, this would indicate some more intimate relationships than discussed so far between these two achievements. And it would emphasize the question for possible evolutionary relationships and coordinated diachronic changes in language and music.

Patel and Daniel (2003) compared music from Britain and France and concluded that the rhythm of British and French music differs in similar ways as the rhythm of British and French speech. English clearly shows the stress-timed rhythm that is, according to Table 1 in Section 2.2, characterized by more complex syllables and a more variable tempo, while French rather shows syllable-timed rhythm, including a relatively low nPVI.

A study by Sadakata *et al.* (2004) shows differences between Dutch and Japanese musicians that we assume to be associated with characteristics of speech rhythm, too. Six percussion players from each of the two countries participated in the experiments as subjects. They "were asked to perform several rhythm patterns consisting of two notes whose duration ratios were 1:1, 1:2, 1:3, 1:4, 1:5, 2:1, 3:1, 4:1 and 5:1 respectively." Dutch and Japanese musicians showed some common tendencies but differed in one important respect: Japanese percussion players performed the extreme ratios (1:4, 1:5, as well as 4:1, 5:1) with a smaller duration ratio than the ratio given by the scores. In our context this would mean that they "distorted" the durations of the notes in the sense of a tendency to equal length, which corresponds to the restricted variability of syllable duration in Japanese, due to the extremely high number of syllable per phrase. Even in the "vocalic nPVI", which is only one component of the variability of the whole syllable, Dutch obtains the second highest and Japanese the second lowest value in Table 1[4].

Further evidence for a possible interaction between speech rhythm and musical rhythm comes from ethnomusicology. Kolinski (1959) reports that the average tempo, *i.e.,* the average number of notes per second, shows a considerably variation among different cultures and languages. For instance about 2 to 2.5 notes per second in his sample of North American Indian languages, and 4 to 5 notes per second in Dahomean, an African language. This amounts to a range of 4 to 10 notes within a span of 2 seconds.

Interestingly, Nettl's (1954) analysis of Arapaho songs "shows that most long and stressed tones are accompanied by long or high vowels." And Meyer's (2007) description of whistled languages suggests that the utterances are intelligible for the reason that they simulate the "acoustic cues carrying the prosody in the spoken voice" as well as the "rhythm that is constitutive" for prosody. These results encourage the search for correspondences between vowel systems and musical notes.

4 Vowels and musical intervals

4.1 Parallels in the inventory size?

According to Rakowski (1999, p. 24), great "care should be taken while making direct comparisons between the phonological system of natural language and that of music" because the "phonemes" of the "language of music" are not as discrete as in natural language. And while the number of musical intervals roughly corresponds with Miller's magical number seven, the number of phonemes in languages is much higher.

We chose a different approach in the search for analogies between specific inventories: The syllable is the basic unit of speech, and the sound (the "sonority") of this unit mainly comes from the vowel that occupies the nucleus of almost any syllable. Thus the vowels are particularly relevant for vocal music, and for music in general, if we assume that the whole musical system is preformed by the patterns of singing. Therefore we expected (Fenk-Oczlon & Fenk, 2005) and indeed found a coincidence between the number of musical intervals and the number of vowels (instead of the total of phonemes):

The most simple style of music, says Nettl (2000), uses three or four pitches. But the pentatonic (5-tone) scale is used more widely than any other formation. According to Burns (1999) it is a widespread pattern not only in non-Western cultures but was also very common in ancient Europe and is still alive in some folk music and in children's songs of the Western culture. He underlines that "the present 12-interval Western scale is probably a practical limit. Any division of the octave into intervals smaller than quarter tones is perceptually irrelevant

for melodic information." (Burns, 1999, p. 257). The existence of a higher number of tones – *e.g.,* through "intervals that bisect the distance between the Western chromatic intervals" - is, at least as a standard in the culture in question (the Arab-Persian system), a rather controversial question (Burns, 1999, pp. 217-18). Burns relates this upper limit of a 12-interval scale to our limited channel capacity. In absolute judgement experiments, a decisive component of the subjects' performance is not only their ability to discriminate between the stimuli but also their ability to identify them, *i.e.,* to recall and assign names or numbers. According to Miller (1956), the range of this performance extends, though in many stimulus materials limited to about 7 categories (2.8 bits), from 3 categories (1.6 bits) to 15 categories (3.9 bits). Burns (1999, p. 223) argues that a "channel capacity of 3.5 bits per octave may indeed be a primary factor in the apparent limitation of 12 notes per octave in musical scales."
As to the musical inventory we may tentatively note an asymmetric frequency distribution that starts at a lower limit of three and ends at an upper limit of twelve elements, showing a prominent peak at five elements. But is the neighbouring 6-tone scale less frequent than the 7-tone scale? The information available is restricted to folk music and music of nonliterate cultures but is nonetheless contradictory: According to the Encyclopaedia Britannica (2006), in such music the hexatonic (6-tone) scales appear rather rarely as compared with the heptatonic scales. Such descriptions insinuate a second but lower peak at seven elements. Nettl (1956, p. 60), however, claims that in "primitive" musical cultures the heptatonic scale is rarer than the hexatonic scale.
In the vowel inventory, the situation is similar. Crothers (1978) found an inventory of 3 - 7 vowels in 80 % of 209 languages investigated, and most of the languages of this sample had 5 vowels. Only 6 languages had more than 9 basic vowel qualities: 12 vowels, which was the maximum, in Pacoh, 11 in French, and 10 in 4 other languages. According to Ladefoged (2001, p. 25) many "languages use just five vowels that can be represented by the letters 'a, e, i, o, u'", and "far more languages have five or seven vowels than have four or six." (p. 35). But as to the frequencies of six versus seven elements we meet again different information: Most of the languages in Maddieson's (2005, p. 14) sample of 563 languages and in Crother's (1978) sample of 209 languages show an inventory of five vowels, but in both studies the next most frequent inventory size is - unlike in Ladefoged - six vowel qualities. And the upper limit? Ladefoged claims a higher maximum than Crothers: 14 or 15 different vowels in General American English (p. 26) and 20 different vowels in the "form of British English used by national newscasters ('BBC English')" (p. 28). The higher maximum in Ladefoged as compared to Crothers is probably due to his rather "liberal" classification system per se and

due to the fact that he also considers inter-individual differences within a language community.

4.2 The colour of vowels in Japanese *shoga* and Alpine yodellers

To compare the whole inventory of musical intervals with only a specific part of the phonemic inventory seems to make sense for several reasons. The vowels represent, as already mentioned, those elements of the phonemic inventory which are of the highest relevance for vocal music, and vocal music might have played a pivotal role in the coordination of language and music. A second reason is the fact that vowels are those elements of the phonemic inventory which differ from each other, in contrast to consonants and similar to musical notes, mainly or even exclusively in two respects: Massively in the frequency of the overtones, and slightly but significantly in their intrinsic pitch, *i.e.,* in their fundamental frequency F0: All other things being equal, high vowels such as [i] and [u] have a higher intrinsic pitch than low vowels such as [a] or [æ]. Whalen and Levitt (1995) could observe this effect in their whole sample of 31 languages. The intrinsic pitch of vowels also shows in tone languages (*e.g.,* Zee, 1980) and in babbling (Whalen *et al.,*1995).

But as to the colour of speech, the frequency of the second formant (F2) seems to be even more important than the frequency of any other formant, as was already suspected by Hockett (1954). A study by Hughes (2000) not only corroborates Hockett's assumption but moreover the view of the vowel system as a link between speech on the one hand and vocal as well as instrumental music on the other: He argues that the "largely subliminal awareness" of the acoustic-phonetic features or the colour of vowels leads in almost any music culture to their use as a mnemonic/iconic system for transmitting or representing melodies. (Among the very few exceptions are the *do, re, mi etc.*, where originally a poetic text was used as an anchor, or the English naming of pitches as *a, b, c etc.,*) For each successive pair of syllables he makes an entry in a matrix showing whether the associated melody pitches ascend, descend, or stay the same. Using this method he found in *shoga,* the Japanese mnemonic system, that "the vowels must be correlated with melodic direction in close correspondence" to their second formant (F2) ordering. For example: If a syllable containing an **i**, which represents the highest position in this hierarchy of the frequencies of F2 (**i, a, o, u**), is followed by a syllable containing an **o**, "the melody at that point descended in 34 of 35 cases in our sample." (Hughes, 2000, pp. 101-02)

Having in mind the sound and vowel patterns of some Alpine yodellers and the freedom of yodellers in combining lexically meaningless syllables with successions of notes, we assumed that this type of music is composed according to the very same principle. Thus we studied all of the monophonic yodellers (n = 15) in Pommer's collection from 1893. The results make inferential statistics unnecessary: In 118 out of 121 cases of a syllable containing an [o] following a syllable containing an [i] the melody descended; in one case the melody raised, in two cases the pitches stayed at the same. And in the 133 instances of an [o]→[i] succession the melody ascended with only one exception (equal pitch). Instances of an [i] followed or preceded by other vowels were less frequent but offered the same picture: one exception in 44 cases of an [i]→[a] succession, and no exceptions in 10 cases of [a]→[i] and in 6 cases of [u]→[i]. It seems that this iconic principle can also be found in the yodelling-like refrains of "Gstanzln" (Austrian-Bavarian mocking songs), and these songs remind of Nettl's (1956:22) description of many archaic styles using meaningless syllables as partial song texts.

5 A Synopsis: Characteristic Values in Language, Music, and Cognition

Table 2[5] comprises basic characteristic values of language and music and assigns them to more general regularities known from behavioural sciences and general/cognitive psychology.

Correspondences in cognition	**LANGUAGE**	**Correspondences in music**
	• duration of syllables: ~ 200 to ~ 400 msec	• n of pulses per min: ~ 30 to ~ 300 (i.e., a minimal duration of ~200 msec per pulse
• duration of "action units" (Schleidt 1992) and of memory spans (psycho-logical present in Fraisse 1957; Baddeley's phonological loop model): roughly 2 sec	• duration of intonation units (e.g. Chafe 1987; Määttä 1993): roughly 2 sec ---------- therefore: a "maximum" of ~10 syllables per rhythmic unit (intonation unit, clause)	• duration of musical phrases corresponds (Parncutt & Pascall 2002) to Fraisse's psychological present: roughly 2 sec ---------- therefore: a "maximum" of ~10 pulses per rhythmic unit (musical phrase)
	cross-linguistic results by the authors:	
• Miller's (1956) magical number seven, plus or minus two	• 5 – 10 (a mean of 6.4) syllables per clause	• 6 to 10 notes in 75 % of phrases in the Essen Folksong Collection (Huron, 1996) and in the Ottmann collection (Temperley, 2001)
• the more single units per action unit, the less time per single movement (Schleidt, 1992)	• the more syllables per clause, the fewer phonemes per syllable	• in phrases containing many notes, the notes are usually very fast (Temperley, 2001)
	• most languages have 5 vowels (Crothers, 1978); different authors claim either 6 or 7 vowels as the next most frequent inventory.	• pentatonic (5-tone) scales are used more widely than any other formation. It is, depending on the source, either followed by the hexatonic (6-tone) scales or by the heptatonic (7-tone) scales.
• a channel capacity of 3.5 bits	• a minimum of 3 and a maximum of 12 vowels according to Crothers (1978)	• 3 tones seem to mark a lower limit (Nettl, 2000), and the chromatic (12-tone) scale an upper limit (Burns, 1999)

So far we have referred to more or less "universal" cognitive constraints in order to explain more or less "universal" parallels between language and music. But if one assumes a more intimate connection between language and music, one should consider evolutionary aspects, too. Did language evolve from music, or music from language, or both language and music from half-musical utterances?

6 Evolutionary perspectives and the role of singing

Many parallels between language and music can be explained by perceptual and cognitive mechanisms involved in both speech and music. For both we need for instance something like an auditory working memory that programs and controls one's own performance, and the increasing complexity of language may have stimulated the evolution of our auditory working memory in a co-evolutionary process (Fenk & Fenk-Oczlon, 2007). Constraints of such mechanisms will be effective in the sense of constraints on the cross-cultural variation of language and music and on the evolutionary pathways of these achievements. But how can we imagine these evolutionary pathways?

One may consider the possibility that both these communication systems originated relatively independently of each other and that they converged later in a process of co-evolution (1). The assumption of a later co-evolutionary process is also compatible with the possibility that music evolved from language (2) or language from (vocal) music (3), or that both evolved from a common precursor (4). Brown (2000), after considering these possibilities, favours the latter and calls it the "musilanguage model".

We understand and use the terms *vocal music* and *singing* in a rather broad sense. *Singing* is neither restricted to something that has to be learned, but is also applicable to the singing of *e.g.,* gibbons (see below). Nor is it restricted to singing meaningful words and texts. This corresponds for instance with Nettl's (1954, p. 192) description of certain Arapaho songs "which have only meaningless syllables". "The world's simplest style", says the ethnomusicologist Nettl, "consists of songs that have a short phrase repeated several or many times, with minor variations, using three or four pitches within a range of a fifth." (Nettl, 2000, p. 469) This is a widespread kind of music. Nettl mentions as examples songs of the Vedda in Sri Lanka, songs of the Yahi tribe in India, music of certain Pacific islands, and "children's ditties of European and other societies, as well as certain pre-Christian ritual songs

preserved in European folk cultures". (Nettl 2000, p. 469) He considers (on the same page) that this archaic style is "what the earliest music of humans was like."

The idea of "music as protolanguage" (3) has many proponents, among them Charles Darwin:

> As we have every reason to suppose that articulate speech is one of the latest, as it certainly is the highest, of the arts acquired by man, and as the instinctive power of producing musical notes and rhythms is developed low down in the animal series, it would be altogether opposed to the principle of evolution, if we were to admit that man's musical capacity has been developed from the tones used in impassioned speech. We must suppose that the rhythms and cadences of oratory are derived from previously developed musical powers. (Darwin, 1871, p. 12)

In order to illustrate that argument, Darwin continues with a vivid description of the singing gibbons. But the singing of the gibbons is not ***singing*** in the sense of *e.g.*, Fitch (2006, p. 194). His main objection is that the acoustic structure of these "songs" develops "reliably in the absence of experience".

Apart from the question whether gibbons really sing, and whether innate singing should in fact be excluded from ***singing***, we think that the thesis of singing as preceding language can be maintained even if one does not look at relatives like the gibbons or look back at very "early progenitors" of man. Skoyles (2000, p. 2) develops a relevant string of arguments: In man, singing requires "(i) the capacity to produce and learn repetitive patterns, and (ii) the thoracic control of expirations to enable long sequences of different tones and articulations made upon a single out breath." Most authors exploring the evolution of breath control and related vocal changes link them to speech. Skoyles, however, argues that "these changes could have evolved first to enable singing, and only then by the addition of vocabulary and syntax became used for speech." (p. 1) One of his arguments comes from ontogenetic evolution:

> The perception of intonation is used (together with distribution regularities in speech sounds) to segment out word boundaries within speech and so enable words to be identified and acquired […] Intonation similarly segments phrase boundaries and thus aids the acquisition of syntax […] Thus, while song could evolve before speech, speech could not have on developmental grounds been acquired without the earlier existence of song. (Skoyles, 2000, pp. 2-3)

Jespersen obviously took both possibilities – language as a descendent of music (3) and both language and music as descendents of holistic half-musical utterances (4) – into consideration: "Language originated as play, and the organs of speech were first trained in the singing sport of idle hours." (Jespersen, 1922, p. 433; quotation from Skoyles 2000, p. 1) This statement by Jespersen is in line rather with position (3), another one, cited in Mithen (2005, p. 2), rather with position (4): Language "began with half-musical unanalysed expressions for individual beings and events." (Jespersen, 1895). This latter statement anticipates the contemporary idea of "musilanguage" (Brown, 2000) as the common roots of language and music.

Regarding the nature of protolanguage, Mithen (2005, p. 3) compares compositional theories as proposed by Bickerton (2000) with holistic theories as proposed by Wray (*e.g.,* 1998). Both the ideas of "half-musical" utterances (Jespersen) and those of "musilanguage" conform more to holistic than to compositional theories. Theories suggesting one precursor become particularly convincing[6] in view of the communication between mother and infant that everyone experiences from changing perspectives in his everyday life.

Trevarthen (1999) focuses on the rhythmic patterns in the mother-infant communication. Again the phrase appears as the central unit and the syllable as its relevant element. But in mother-infant communication the tempo slows down to only two or three syllables per second:

> The timing of this face-to-face play is that of a friendly adult chat or discussion […] Each infant utterance, with its vocalisation, lip-and-tongue movements and hand gestures, lasts about 2 or 3 seconds, about the time an adult takes to say a phrase or a short sentence. The individual coos last only about a third to half a second, comparable with a syllable. (Trevarthen, 1999, p. 176)[7]

In Stern (2001, pp. 147-48) such intervals exceeding 2 sec seem to be rather a sum of vocalization plus switching pause. Referring to Stern and particularly to Trevarthen (1999) Cross suggests "that human infants interact with their caregivers in ways that appear to be 'proto-musical' " (Cross 2003, p. 79) and McMullan & Saffran (2004) assume similar developmental underpinnings in language and music and that modularity of these domains is emergent rather than present at the beginning of life.

Let us summarize: Singing without words – as in the songs of animals, in "animal-like songs", in many utterances of our infants, in some yodelers, in certain Arapaho songs (Nettl, 1954) – is cognitively less demanding than the use of a more or less arbitrary code. And it will function well without and before the singer has become familiar with such a code, as illustrated by the examples above. This reduces the plausibility of language as a precursor of music. For a closer inspection there remain the two possibilities already favored in Jespersen (1895): Music either as precursor of language or both language and music as descendents of "half-musical" utterances.

But with respect to tonal modulation, singing is more demanding than speech. For this and some other reasons it is more plausible that singing "prepared" the vocal tract for speech than the other way round. Thus we are sympathetic with a complement to the musilanguage model, *i.e.,* with the assumption "that tonality was the ancestral state of language" (Brown, 2000, p. 281), and with Morley's (2003) description of a "progressively increasingly complex proto-language based on tone-dominated affective social utterances."(p. 149). Let us take the most archaic singing described in Nettl (2000, p. 469) as a model: Simple musical phrases repeated, modified, and repeated again, are an appropriate production system as well as an ideal carrier current system for transmitting symbols, *i.e.,* elements of a coding system associated with particular meanings by virtue of their rule-based use. The discovery of such phrasal segments as a matrix for words and one-word sentences, for meaningful phrases and formulaic speech, must have initiated a selective pressure on the evolution of a referential system and maybe also a co-evolutionary, mutual stimulation of progress in language and cognition.

This model not only allows for the correspondences between language and music in the repertoires, in phrase length and rhythmic structure of phrases. It also sheds some light on the role of the vowel system: Sound and colour are inherent properties of vowels, and in speech – the specialist for referential meaning (Brown, 2000) – this sound diminished as compared with a less specialized, tone-dominated precursor of song and speech.

7 Discussion

From the evolutionary models discussed we would prefer the idea of tone-dominated affective utterances as the basis of both vocal music and a protolanguage decreasing in tone but taking on a more articulate syllable structure required as a carrier for the semantic units we call "words" and "propositions". This idea well fits the correspondences found between language and music regarding their rhythmic organization and their segmental inventories.

On the other hand we think that these parallels might be explicable to some degree in a more parsimonious way, *i.e.,* without recourse to phylogenetic theories.

- Constraints and limits of our cognitive system are effective in both achievements. The ideal size of the "packages" that our cognitive apparatus can simultaneously handle seems to be in the region of 5 rather complex or 10 very simple syllables within an intonation unit and in a region of 6 to 11 notes within a musical phrase (Huron, 1996, Temperley, 2001).

In terms of a multi-store model the limits relevant for these packages might be attributed to working memory constraints, while the restricted inventories – preferably 5 to 7 vowels in language and 5- to 7- tone scales in music – rather reflect long-term memory constraints. But even within such a model one may consider a transition of working memory constraints into long-term memory (Fenk-Oczlon & Fenk, 2000).

- The most original form of music – most original in the ontogenetic and maybe also in the phylogenetic evolution and prehistoric development - is probably singing (Nettl, 2000), and the most common form of music is still singing or singing accompanied by instruments. Thus it is plausible to assume that singing shaped the form of instrumental music. Talking as well as singing comes about in intonation units. If intonation units are regarded as a special case of action units (Fenk-Oczlon & Fenk, 2002), one has to assume that any determinant of intonation units will be reflected in language as well as in music. These determinants are the "clausal structure" of the breath cycle, the (coordinated) "clausal structure" of cognitive processes that becomes apparent when *e.g.,* focusing a Necker Cube (*cf.,* Pöppel, 1985), and, last but not least, the cognitive activities in programming and controlling the tempo changes of breathing and the shape and articulation of sound.

An intimate coordination between action units and perception units might have played a pivotal role in the emergence of language and in the co-evolution of language, or "musilanguage", and cognition. Such a precise sensorimotor coordination should also show in other sorts of highly skilled performance – especially in activities requiring, in addition, interindividual coordination as is the case in some common efforts in work, like those synchronized with the help of shanties, or in dance, which is most commonly accompanied by music, and in music as such.

Acknowledgements

We would like to thank the guest editors Oliver Vitouch and Olivia Ladinig for their support and Andrzej Rakowski as well as two further anonymous reviewers for their detailed and helpful comments.

Notes

[1] Quite interesting in this context: According to Ramus *et al*., (2000), human newborns and Tamarin Monkeys can discriminate between sentences of exactly those two languages marking the endpoints of our variable syllable complexity.

[2] Thai is one of the languages not yet included in our regressional analyses of 34 languages. A second and more complex statistical analysis will follow in the near future with a meanwhile extended sample of about 60 languages.

[3] It would be interesting but would go beyond the scope of this paper to correlate our complexity measures with the measures suggested by Ramus *et al*. (1999) as the most appropriate ones for distinguishing between stress-timed and syllable-timed languages: percentage of vowels (%V) and standard deviation of consonantal intervals (ΔC). We would suspect high correlations for two reasons: If we take for granted that (almost) any syllable has one and only one vowel, then an utterance's n of syllables will be (almost) identical with its n of vowels. And ΔC depends to some degree on syllable complexity: a high number of elements is a presupposition for high variability.

[4] In any language almost any syllable contains one vowel, a monophthong or diphthong, additional complexity and duration is mainly conveyed by additional consonants.

[5] The rather sweeping use of the terms "maximum" and "upper limit" in Table 1 and in different contexts and authors would, of course, deserve a more differentiated analysis, e.g., with respect to the object and nature of the relevant data. ("Performance" of populations or individuals, of persons or of communication systems? Single data or aggregated data?) In our cross-linguistic studies these terms refer to aggregated data: 10 syllables per simple declarative sentence is neither a theoretical nor an empirical maximum but the highest MEAN VALUE we found in a standardized method in our hitherto sample of 34 languages.

[6] In this context, Mithen (2005) reminds the reader of Haeckel's (1866) recapitulation theory which suggests that ontogenetic evolution recapitulates phylogenetic evolution. In contemporary biology it is viewed rather as a heuristic principle than a biological rule. And it will become rather fuzzy when applied to (the organic basis of) young achievements such as tool making and *language* or *singing* in the narrow sense of the word.

[7] The phrase length (2-3 sec) reported by Trevarthen (1999) is in line with a neural, low frequency mechanism (about 3 sec) that "binds individual musical elements into rhythmical *Gestalts*." (Wittmann & Pöppel 1999, p. 13). The boundaries of such rhythmic and tonal patterns also shows in event related potentials (ERPs). A Contingent Negative Variation (CNV) initiated by the onset of a sentence and ending with a positive shift at its end could already be observed in the N 400 experiments by Kutas & Hillyard (1980). A systematic investigation by Steinhauer, Alter & Friederici (1999) revealed the coincidence of a "closure positive shift" (CPS) with the end of intonational phrases within longer sentences, *e.g*., one CPS in the first one and two CPSs in the second one of the following conditions: "*Peter verspricht Anna zu arbeiten* /CPS/ *und das Büro zu putzen*."; "*Peter verspricht* /CPS1/ *Anna zu entlasten* /CPS2/ *und das Büro zu putzen*." A positive shift, though not identical with language-CPS in distribution, latency and duration, is also associated with phrase boundaries in music (Knösche *et al.* 2003). In this study the generators of the CPS were found to be localized in bilateral *planum temporale*.

References

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.

Bickerton, D. (2000). How protolanguage became language. In C. Knight, M. Studdert-Kennedy, & J. Hurford (Eds.), *The evolutionary emergence of language* (pp. 264-84). Cambridge: Cambridge University Press.

Brown, S. (2000). The "musilanguage" model of human evolution. In N.L. Wallin, B. Merker & S. Brown (Eds.), *The origins of music* (pp.271-300). Cambridge, MA: The MIT Press.

Burns E. M. (1999). Intervals, scales, and tuning. In D. Deutsch (Ed.), *The psychology of music* (pp. 215-64). 2nd ed. San Diego: Academic Press.

Chafe, W. (1987). Cognitive constraints on information flow. In R. S. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 21-51). Amsterdam: John Benjamins.

Chu, M., & Feng, Y. (2001). Study on factors influencing durations of syllables in Mandarin. *Proceedings of Eurospeech* (pp. 927–30). Aalborg.

Cross, I. (2003). Music and evolution: causes and consequences. *Contemporary Music Review*, *22*, 79-89.

Cowan, N. (2001). The magical number 4 in short-term memory. A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-114.

Crothers, J. (1978). Typology and universals of vowel systems. In J. H. Greenberg (Ed.), *Universals of human language* (pp. 93-152). Stanford: Stanford University Press.

Cummins, F. (2002). Speech rhythm and rhythmic taxonomy. In *Proceedings of Prosody* (pp. 121-26), Aix en Provence.

Darwin, Ch. (1871). *The descent of man and selection in relation to sex*. http://www.darwin-literature.com/The_Descent_Of_Man/21.html. 8/24/2004.

Encyclopaedia Britannica (2006) http://www.britannica.com/eb/article-64512/scale, 02.10.2006.

Fenk, A., & Fenk-Oczlon, G. (1993). Menzerath's law and the constant flow of linguistic information. In R. Köhler & B.B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 11-31). Dordrecht/Boston/London: Kluwer Academic Publishers.

Fenk, A., & Fenk-Oczlon, G. (2007). Inference and reference in language evolution. In S. Vasniadou, D. Kayser, A. Protopapas (Eds.), *Proceedings of EuroCogSci07* (p. 889). Hove: Lawrence Erlbaum Associates.

Fenk-Oczlon, G. (1983). *Bedeutungseinheiten und sprachliche Segmentierung. Eine sprachvergleichende Untersuchung über kognitive Determinanten der Kernsatzlänge*. Tübingen: Gunter Narr.

Fenk-Oczlon, G., & Fenk, A. (1985). The mean length of propositions is 7 plus minus 2 syllables – but the position of languages within this range is not accidental. In G. d'Ydewalle (Ed.), *Proceedings of the XXIII International Congress of Psychology: Selected/Revised Papers*, *Vol. 2: Cognition, information, and motivation* (pp.355-59). Amsterdam: Elsevier Science Publishers.

Fenk-Oczlon, G., & Fenk, A. (1999). Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology*, *3-2*, 151-77.

Fenk-Oczlon, G., & Fenk, A. (2000). The magical number seven in language and cognition: empirical evidence and prospects of future research. *Papiere zur Linguistik*, *62/63*, 3-14.

Fenk-Oczlon, G., & Fenk, A. (2002). The clausal structure of linguistic and pre-linguistic behavior. In T. Givón & B.F. Malle (Eds.), *The evolution of language out of pre-language* (pp.215-29). Amsterdam: John Benjamins.

Fenk-Oczlon, G., & Fenk, A. (2005). Crosslinguistic correlations between size of syllables, number of cases, and adposition order. In G. Fenk-Oczlon & C. Winkler (Eds.) *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler* (pp. 75-86). Tübingen: Gunter Narr.

Fenk-Oczlon, G., & Fenk, A. (2005). Cognitive constraints on the organization of language and music. In B.G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (p. 2476). Mahwah, NJ: Erlbaum.

Fenk-Oczlon, G., & Fenk, A. (2006). Speech rhythm and speech rate in crosslinguistic comparison. In R. Sun & N. Miyake (Eds.) *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (p. 2480). Mahwah, NJ: Erlbaum.

Fitch, W. T. ((2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100 (1)*, 173-215.

Fraisse, P. (1957). *Psychologie du temps*. Paris: Presses Universitaires de France.

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.), *The psychology of music* (pp. 149-80). New York: Academic Press.

Gigler, G. (in preparation). *Intonationseinheiten im Kaerntnerischen*. Doctoral dissertation, University of Klagenfurt.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.) *Laboratory Phonology 7* (pp. 515-46). Berlin: Mouton de Gruyter.

Greenberg, J. H. (1968). *Anthropological linguistics: An introduction*. New York: Random House.

Hockett, C. F. (1955). *A manual of phonology*. Baltimore: Waverly Press.

Hughes D.W. (2000). No nonsense: the logic and power of acoustic-iconic mnemonic systems. *British Journal of Ethnomusicology, 9/ii,* 95-122.

Huron, D. (1996). The melodic arch in Western folksongs. *Computing in Musicology, 10,* 3-23.

Jespersen, O. (1922). *Language.* London: Allen.

Jespersen, O. (1983 [1895]). Progress in language. *Amsterdam Classics in Linguistics 17*. Amsterdam: John Benjamins.

Kegel, G. (1990). Sprach- und Zeitverarbeitung bei sprachauffälligen und sprachunauffälligen Kindern. In G. Kegel, T. Arnhold, K. Dahlmeier, G. Schmid & B. Tischer (Eds.), *Sprechwissenschaft und Psycholinguistik, 4*. Opladen: Westdeutscher Verlag.

Knösche, T. R., Haueisen, J., Neuhaus, C., Maess, B., Alter, K., Friederici, A. D., & Witte, O.W. (2003). An electrophysiological marker for phrasing in music. Paper presented at the *HFSP Workshop on Music and Language Processing*, Sydney, April 2003.

Kolinski, M. (1959). The evaluation of tempo. *Society of Ethnomusicology, 3*, 45-57.

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology, 11*, 99-116.

Ladefoged, P. (2001). *Vowels and consonants*. Malden: Blackwell.

Lavoie P., & Grondin, S. (2004). Information processing limitations as revealed by temporal discrimination. *Brain and Cognition, 54*, 198-200.

Luangthongkum, F. (1977). *Rhythm in standard Thai*. Unpublished Ph.D. thesis. University of Edingburgh.

Määttä, T. (1993). Prosodic structure indices in a folk tale. *Phonum, 2*, 107-20.

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, 79, 487-509.

McMullen, E., & Saffran, J. R. (2004). Music and language: A developmental comparison. *Music Perception, 21*, 289-311.

Merker, B. (2002). Music: The missing Humboldt system. *Musicae Scientiae, 6*, 3-21.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Mithen, S. (2005). *The singing Neanderthals*. London: Weidenfeld & Nicolson.

Meyer, J. (2007). What does the typology of whistled forms of language teach us about prosody? In Book of Abstracts (p. 173), *7th International Conference of the Association for Linguistic Typology*, Paris, September 25-28.

Morley, I. (2003). *The evolutionary origins and archaeology of music*. Doctoral Dissertation, University of Cambridge (electronic edition Jan. 2006).

Nettl, B. (1954). Text-music relationships in Arapaho songs. *Southwestern Journal of Anthropology, 10*, 192-99.

Nettl, B. (1956). *Music in primitive culture*. Cambridge: Harvard University Press.

Nettl, B. (2000). An ethnomusicologist contemplates universals in musical sound and musical culture. In N.L. Wallin, B. Merker and S. Brown (Eds.), *The origins of music* (pp.463 - 72). Cambridge, MA: The MIT Press.

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition*, *87*, B35-45.

Parncutt, R., & Drake, C. (2001). Psychology: Rhythm. In S. Sadie (Ed.), *New Grove Dictionary of Music and Musicians*, *20*. London, 535-38, 542-55.

Pike, K. L. (1945). *The intonation of American English*. Ann Arbor: University of Michigan Press.

Parncutt, R., & Pascall, R. (2002). Middle-out music analysis and its psychological basis. In C. Stevens, D. Burnham, G. McPherson, E. Schubert & J. Renwick (Eds.), *Proceedings of the 7th International Conference on Music Perception and Cognition.* Adelaide: Causal Productions.

Pommer, J. (1893). *252 Jodler und Juchezer*. Wien: Rebay & Robitschek.

Rakowski, A. (1999). Perceptual dimensions of pitch and their appearance in the phonological system of music. *Musicae Scientiae, 3, 1*, 23-29.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm. In B. Bernard & I. Marlien (Eds.) *Proceedings Speech Prosody 2002*. Aix-en-Provence.

Ramus, F., Nespor, M., & Mehler J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*, 265-92.

Ramus, F., Hauser, M.D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by Cotton-top Tamarin Monkeys. *Science*, *288*, 349-51.

Sadakata, M., Ohgushi, K., & Desain, P. (2004). A cross-cultural comparison study of the production of simple rhythm patterns. *Psychology of Music*, *32*, 389-403.

Schleidt, M. (1992). Universeller Zeittakt im Wahrnehmen, Erleben und Verhalten. *Spektrum der Wissenschaft*, Dezember, 11-115.

Schleidt, M., & Kien, J. (1997). Segmentation in behavior and what it can tell us about brain function. *Human Nature, 8*, 77-111.

Sloboda, J.A. (1982). Music Performance. In D. Deutsch (Ed.) *The psychology of music.* (pp. 479-96) New York: Academic Press.

Skoyles, J.R. (2000). The singing origin theory of speech. *Abstracts 3rd Conference "The Evolution of Language"* http://www.infres.enst.fr/confs/evolang/actes/_actes65.html

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. nature *neuroscience*, *2/2*, 191-96.

Stern, D. N. (2001). Face to face play: its temporal structure as predictor of socioaffective development. *Monographs of the Society for Research in Child Development*, *66*, 144-49.

Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge/Mass, London: MIT Press.

Trevarthen, C. (1999). Musicality and the intrinsic motive pulse: Evidence from human psychobiology and infant communication. *Musicae Scientiae, Special Issue (1999-2000)*, 155-215.

Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics*, *10*, 193-218.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics, 23,* 349-66.

Whalen, D. H., Levitt, A. G., Hsiao, P.-L., & Smorodinsky, I. (1995). Intrinsic F0 of vowels in the babbling of 6-, 9- and 12-month-old French-and English-learning infants. *Journal of the Acoustical Society of America, 97,* 2533-39.

Wittmann, M., & Pöppel, E. (1999). Temporal mechanisms of the brain as fundamentals of communication – with special reference to music perception and performance. *Musicae Scientiae, Special Issue (1999 - 2000)*, 7-11.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication, 18*, 47-67.

Zee, E. (1980) Tone and vowel quality. *Journal of Phonetics, 8,* 247-58.