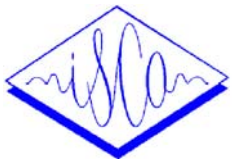

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

Cerebral Strategies in the Segmentation and Interpretation of Speech

Ulrike Toepel & Kai Alter

Max-Planck-Institute of Cognitive Neuroscience
Stephanstr. 1a, 04103 Leipzig, Germany
toepel/alter@cns.mpg.de

Abstract

The segmentation of the acoustic speech signal is a fundamental for the processing of spoken language. The paper at hand provides a survey of studies conducted in our lab concerning the detection of segmentation cues in the speech signal and associated perception of prosodic boundaries.
The first two studies presented here employ the methodology of Event-Related Potentials (ERP) to study online electrophysiological responses to acoustic stimuli varying in syntactic and prosodic constituency, as well as in segmental content.
By the first study an ERP shift was identified correlated with the perception of major intonational boundaries which was termed the Closure Positive Shift (CPS).
The second study was especially concerned with listener's abilities in speech segmentation, given the exclusive presence of prosodic cues.
A third experiment reviewed here employs functional Magnetic Resonance Imaging (fMRI), an investigation method based on hemodynamic brain responses.
ERP and fMRI are complementary methodologies: while ERPs provide an accurate measure of temporal aspects of processing, fMRI methodology is particularly well suited to localize such processes in the brain.

10.21437/SpeechProsody.2002-10

1. Introduction

It is increasingly accepted that processing of linguistic information, in terms of semantics, syntax and phonology, is always associated with intrinsic prosodic cues to its meaning. Interestingly, these cues do not only seem to be provided by speech and hence explicit acoustic parameters but also by an implicit prosody of visually presented language [4, 18].
But also for auditory language processing the role of prosodic information in linguistic interpretation is still a matter of controversy. Although a great number of studies have focused on the auditory correlates of prosodic structure, including duration, pitch and intensity [3], surprisingly little research has yet been done on the question how the human brain processes these cues to linguistic structure *online*.
The current investigation method with the highest temporal solution is the methodology of event-related potentials (ERP), in which an electroencephalogram (EEG) signal is measured and mapped onto an event. The advantage of the ERP is the computing of a *time-locked* cerebral response to certain sensory stimulation, i.e. also by linguistic means.
ERPs have been used successfully for more than a decade to investigate semantic and syntactic processing of visual presented stimuli [6, 8].
Recently technological innovation has yielded sufficiently accurate identification of prosodic parameters in the speech signal to allow investigation of electrophysiological responses to prosodic processing.
In an initial investigation, presented in section 2, Steinhauer, Alter and Friederici [19] identified the Closure Positive Shift (CPS), an ERP component correlated with the perception of major prosodic boundaries. In addition to being the first electrophysiological correlate of linguistic prosody, it was the first ERP component that could be elicited during normal speech processing without the employment of expectancy or structural violations.
The following line of argumentation is supposed to clarify the importance of prosodic as well as segmental-linguistic cues for evoking the CPS in the ERPs and provides data on the interpretation of delexicalized speech input by the human brain. Although it has been shown that the CPS can under certain circumstances also be evoked during visual language processing [18] and already exists during language acquisition [9], it is not yet really clear whether the exclusive existence of prosodic information in the speech stream is sufficient for *eliciting* the Closure Positive Shift. In this regard, contrary findings will be reported.
Additionally, a fMRI experiment comparing the general processing of normal and delexicalized speech will be reviewed here, which also provides insights into the cerebral structures employed in prosody perception. It presents evidence for hemispheric variability in the perception of pure linguistic vs. prosodic cues (for a review, see [1]).

2. The Closure Positive Shift as an indicator of the perception of prosodic domains [19]

ERP experiments using auditory presentation have primarily been concerned with replicating electrophysiological components already identified in visual language processing.
Spoken language differs however from written language in carrying overt prosodic cues, which have been shown to function to resolve syntactic and semantic ambiguities (for a review, see [3]).
Astonishingly, no study so far had addressed the online perception of utterance-internal phrase markers.
For this purpose, a sentence corpus was developed varying phrasal complexity.

A1	[Peter]$_{NP1}$ [verspricht]$_{VP1}$ [Anna]$_{NP2}$ [zu arbeiten]$_{VP2}$ [und das Büro zu putzen]
	Peter promises Anna to work and to clean the office. (literally)
B1	[Peter]$_{NP1}$ [verspricht]$_{VP1}$ [Anna]$_{NP2}$ [zu entlasten]$_{VP2}$ [und das Büro zu putzen]
	Peter promises to support Anna and to clean the office. (literally)

In both examples the first verb phrase 'verspricht/ promises' is ambiguous with regard to its complexity. The second verb phrase is intransitive in condition A1 and transitive in condition B1. So, the second NP 'Anna' becomes the direct object of VP1 in condition A1 but the object of the infinitive VP2 in condition B1.
In visual language processing this attachment ambiguity would lead to processing difficulties in readers but due to the immediate availability of prosodic cues in speech listeners should not even notice the existence of an ambiguity.
To prove prosodic variability between the conditions due to their differing syntactic structure exhaustive acoustic analyses were conducted.

2.1. Speech Signals

48 sentences were produced by a female speaker of standard German and recorded in a soundproof chamber. The stimuli were digitized (44.1 kHz/16 bit sampling rate) and a range of acoustic analyses were carried out (measurement of constituent and pause durations, tracking of the fundamental frequency [F0]). These were then subjected to statistical analyses.
Results showed that the speaker had indeed produced a differing intonational phrasing. In the intransitive condition A1 the first Intonational Phrase (IPh) boundary [15] with a high boundary tone, lengthening of the prefinal syllable and a pause is located after the second VP whereas the transitive condition B1 contains an additional IPh boundary after the first VP.

A1	[NP1 VP1 NP2 VP2]$_{IPh1}$ [conjunction]$_{IPh2}$
B1	[NP1 VP1]$_{IPh1}$ [NP2 VP2]$_{IPh2}$ [conjunction]$_{IPh3}$

After the acoustic analyses had been carried out, the stimuli were integrated into a perceptual experiment employing the methodology of ERPs. The aim was to detect potential brain responses to the processing of prosodic boundaries.

2.2. Subjects and Task

Twenty volunteers participated in the study. All of them were right-handed [12] and without hearing or neurological disorders. To keep them attending to the presented auditory input they had to answer a comprehension question in 20 % of the trials. Questions were of the kind 'Does Anna promise to clean the office?'. Altogether, 144 experimental sentences were presented intermixed with 144 filler sentences. The electroencephalogramm (EEG) was continuously recorded from 17 cap-mounted electrodes while subjects were sitting in an electromagnetically shielded chamber. Event-related potentials were then computed from the EEG.

2.3. Electrophysiological responses

ERPs are a transient change of voltage, reflecting a systematic brain activity due to physical events and with implicit on-line

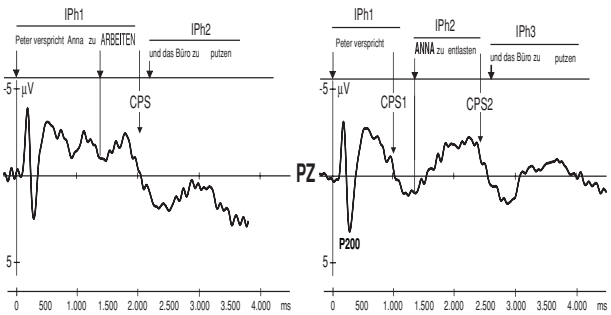

Figure 1: *ERPs to natural speech with differing syntactic and hence prosodic phrasing at the parieto-central electrode. Responses to the intransitive condition (A1) are illustrated in the left diagram, the right panel shows the cerebral reactions to the transitive condition (B1)*

characteristics. In the present study this event is the presence of acoustic correlates of prosodic structure.
In Figure 1 the ERPs of listeners for condition A1 (left panel) and condition B1 (right panel) is illustrated.
In congruence with the acoustic analyses a positive-going waveform is observable in the position of the IPh-boundary after VP2 (negativity is plotted upwards) in both conditions. By contrast, only condition B1 elicits a positivity in the ERP after VP1. This shift corresponds to the existence of an additional IPh-boundary in condition B1 as revealed by the acoustic analyses.
Please note that both conditions do not differ in word order nor employ structural violations. Nevertheless, the sentence conditions have to be processed differently:
In condition B1 the presence of an additional IPh-boundary after VP1 prevents the integration of the following syntactic material into the first verb phrase and prepares an initial attachment of NP2 to the transitive second verb phrase. This parsing preference can only be triggered by prosodic information since up to this point the structural content of the sentences is obviously the same.
This positive-going waveform was termed Closure Positive Shift (CPS) by the authors since it presents an immediate cerebral response to sentence segmentation into prosodic units, i.e. IPhs.

3. The Closure Positive Shift in extended processing conditions

In the previous section, we presented results establishing the CPS as psychophysiological indicator of IPh boundaries [15] in the acoustic speech signal. To further investigate the nature of this brain potential an additional corpus of material was developed varying once the syntactic and hence prosodic constituency of the utterances and their overall segmental content.
Regarding the first demand sentences were derived by inserting either parentheticals (A2) or temporal adjuncts (B2) into the original material. By doing so, we hoped to separate syntactic from prosodic processing cues, as parentheticals are thought to establish a prosodically independent IPh [16]. In this regard, we note that Cooper & Paccia-Cooper [2] have argued that the boundary parameter at the right edge of a parenthesis can be aligned with the right edge of a syntactic constituent, whereas

this is not the case for the left edge.
In contrast, the temporal adjunct was expected to merge its intonation contour as a hierarchically lower phonological phrase part (PPh) into a surrounding higher IPh.
Examples of the material with a parenthetical (A2) and with a temporal adjunct (B2) are given below.

A2	[Peter verspricht Anna]$_{IPh}$ [das weiss Ingo]$_{IPh}$ [zu arbeiten]$_{IPh}$ [und das Büro zu putzen]$_{IPh}$
	Peter promises Anna, Ingo knows that, to work and to clean the office. (literally)
B2	[Peter verspricht Anna]$_{IPh}$ [[am Donnerstag]$_{PPh}$ [zu arbeiten]$_{IPh}$ [und das Büro zu putzen]$_{IPh}$
	Peter promises Anna to work on thursday and to clean the office. (literally)

If the CPS electrophysiological response is reflective of IPh boundary processing alone, then it should occur regardless of a congruent syntactic boundary.
A second question in this study concerns the relevance of additional phonemic content for the occurence of the CPS. All of the former CPS studies had used sentence material containing additional segmental information (i.e. phonemic, syntactic and semantic) in some way.
For exploring the role of pure prosodic information and its influence on the CPS the material was manipulated to exclusively contain prosodic parameters.
The delexicalization procedure adopted for these purposes [17] firstly extracts the pitch marks from the original signal (condition A2 and B2) and then replaces them with three superimposed sinusoidal signals. Regarding the spectral quality of the signal, all frequencies above the third harmonic are being removed. Unvoiced segments of the original signal are being set to zero so that they still reflect the original rhythmic structure. The evolved signal then only comprises the prosodic parameters fundamental frequency, intensity and duration and sounds like "humming" behind a door (condition C2 and D2).
Since the CPS has been shown to be an indicator of prosodic phrasing it was hypothesized that whenever the detection of Intonational Phrase boundaries is exclusively relying on prosodic information in the auditory speech signal, delexicalized speech material should also be sufficient for evoking the CPS.

3.1. Speech signals

The two natural speech conditions with forty-eight sentences each were produced by a female native speaker of standard German and recorded in a soundproof chamber. The digitized speech signals (44.1 kHz/16 bit sampling rate) were separately measured with respect to word and pause durations and fundamental frequency. Thereby the differing prosodic boundary pattern of the conditions (A2 vs. B2) was confirmed. Condition A2 exhibited an additional IPh boundary pattern (IPh) in the time window between 2500-3000ms after sentence onset with a significant lengthening of the prefinal syllable, a high boundary tone and also a significantly longer pause. This time frame can be aligned with the right edge of the inserted parenthesis in condition A2.
For the delexicalization the stimuli were downsampled to 16 kHz due to the requirements of the filtering procedure. In Figure 2 the signal before and after the filtering procedure is being visualized.

3.2. Subjects and Task

Twenty-one volunteers (11 female) without any known hearing or neurological disorders participated in the study. All subjects were right-handed according to the Edinburgh Handedness Inventory [12]. The experiment complied with German legal requirements. To ensure the participants' attention during the experiment they were asked to compare acoustic signals in 20% of the trials. In half of the comparison trials the signals were identical with respect to their intonation contour (once taken from the natural speech conditions and once from the delexicalized ones), whereas in the other 50% the signals were taken from opposite conditions (A $\leftrightarrow$ D; B $\leftrightarrow$ C).

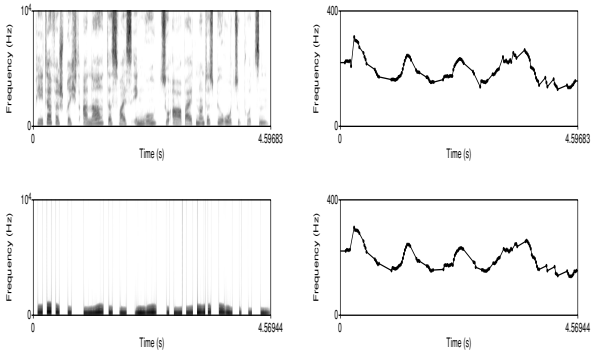

Figure 2: *Illustration of the spectral parameters (left) and intonation contours (right) of the stimuli (condition A2 here as representative) before (upper panel) and after (bottom panel; evolved condition C) the delexicalization procedure.*

3.3. Results

The participants were not able to judge the acoustic signals as being derived from the same or different conditions. They acted on chance level.

3.3.1. Electrophysiological responses in the natural speech conditions

For the statistical analyses nine time-windows (TW) of 500ms each were formed, covering the entire sentence length. This procedure enabled the mapping of an acoustic event onto a perceptual response in the ERPs.
The condition A2 (Figure 3) exhibits a large positive shift in the TW between 2500-3500ms whereas this is not the case for condition B2. Statistical analyses confirmed significant differences between the conditions and showed that those are always strongest at the parieto-central electrode.
The occuring positivity also coincides with the additional intonational phrase boundary of condition A2 as shown in the speech signal section.

3.3.2. Electrophysiological responses in the delexicalized speech conditions

The delexicalized stimulus conditions only comprising prosodic information (C2 and D2) exhibit a large negativity at frontal electrodes when compared to the natural speech conditions (see

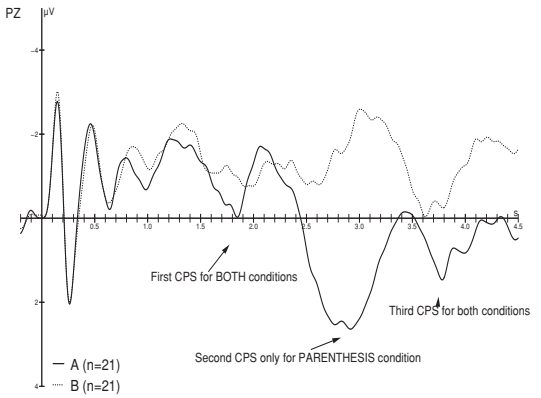

Figure 3: *Electrophysiological responses to only the natural speech conditions (A2 and B2) at the centro-parietal electrode.*

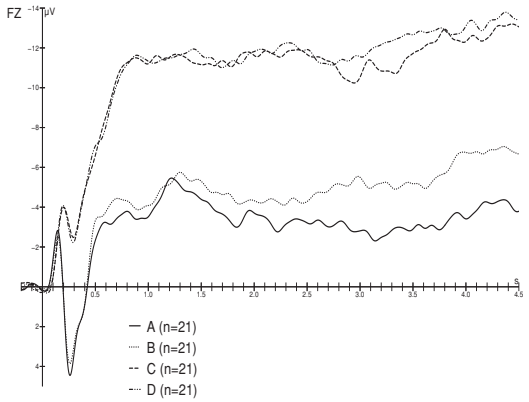

Figure 4: *Electrophysiological responses to natural and delexicalized speech illustrated at the fronto-central electrode.*

Figure 4). Statistic analyses reveal significant overall processing differences between the two classes of stimuli (natural vs. delexicalized) over the entire sentence length (TW 0-4500ms). Examining exclusively the responses for the delexicalized conditions (TWs of 500ms), no significant differences between them are revealed. This is, although the *prosodic* parameters proposed to be responsible for evoking the CPS are the same as in the natural conditions. An additional analysis revealed no processing differences between the left vs. right hemisphere.

4. Hemodynamic responses to natural and delexicalized speech [10]

The study utilized functional Magnetic Resonance Imaging (fMRI) to localize distinct brain areas involved in the processing of linguistic prosody in (again) natural and delexicalized speech.

As opposed to the ERP studies reviewed above, the focus of the current study was no further elaboration of the nature of the CPS. Since the hemodynamic response in fMRI studies has a very rough temporal solution, such a temporally fine-grained phenomenon cannot be investigated by this method. Therefore, this study was more concerned with the overall processing differences between normal and delexicalized speech. Furthermore, the hemispheric contributions to the perception of this differing acoustic input were to be explored.

From a simplified acoustic point of view, natural speech can be divided into two frequency portions - into high or spectral frequencies and into low frequencies. The lowest frequency in the speech signal is called fundamental frequency (F0) and corresponds roughly to the perceptual category of pitch. Higher frequencies can be analyzed as formant frequencies (at least for the sonorant parts of the speech signal). Under this view, prosodic information is related to the low frequency parts of natural speech, e.g. to F0-variation.

An auditory input consists of a more or less well-ordered progression of superimposing frequencies. The perceptual system has to filter out those parameters necessary for comprehending language from the incoming stream of spectro-temporal information.

In the first instance, speech sounds and their combinations, syllables and words, must be segmented. As soon as syllables and word forms are identified, the corresponding word entry in the lexicon can be activated. Thus, the path from hearing to comprehension can be seen as a succession of individual steps of processing.

The listener's system starts off with primary auditory processing; underlying physical events in the acoustics of the speech signal must be combined into more complex units by frequency and time analysis.

This proceeds in an incremental and parallel fashion with, e.g. grammatical analysis, before integration with conceptual and world knowledge takes place.

Whenever information at higher processing levels is not provided, i.e. for grammatical analysis, the brain receives deviant input, it has to compensate for, in order to interpret the stimulus. It was hypothesized that this compensation mechanism results in an increase of local blood supply to cerebral structures, which are not necassarily involved into natural speech processing.

4.1. Material

The material consisted of four sentence conditions with either provided or suppressed prosodic information. The delexicalized material was again derived by applying the filtering procedure mentioned in the ERP study above [17]. Suppression of pitch information was done by flattening the intonation contour of the utterances.

Only the comparison between the natural and the delexicalized speech condition will be reported here. For a more detailed analysis, see the paper of Meyer et al. [10].

4.2. Subjects and Task

All participants (12 right-handed German native speakers from the University of Leipzig) were asked to perform a prosody comparison similar to the design described for the ERP study described above.

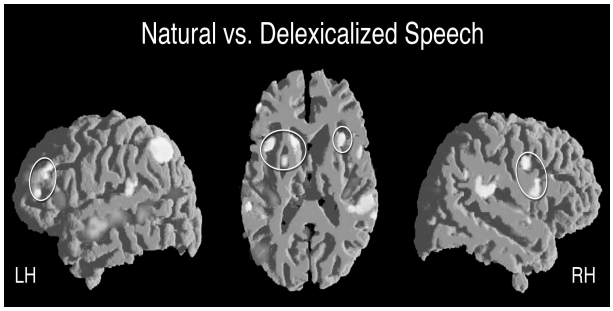

Figure 5: *Additional cerebral responses to delexicalized speech as opposed to natural speech.*

4.3. Results

Generally, the hemodynamic response to *natural speech* was stronger in the left as compared to the right hemisphere. For this condition, also a stronger activation in the supratemporal regions of both hemispheres (with primary auditory cortex and its association areas), in the pars triangularis of the left inferior frontal gyrus (adjacent to Broca's area) and in the subcortical thalamus can be observed.
When *delexicalized speech* had to be processed (see Figure 5), activations in the left and right pars opercularis of the inferior frontal gyrus as well as in other areas of the right hemisphere (posterior Sylvian fissure) or, respectively, of the left hemisphere (middle frontal gyrus and middle cingulate gyrus) increased. Also subcortical activation is observable.
The weaker activation in the left and right supra-temporal region can be interpreted by the fact that the delexicalized material does not contain phonemic information which could contribute to speech understanding.
Concerning the functional lateralization, the processing of delexicalized speech is related to an increase of activation in the posterior part of the right temporal plane. This finding is consistent with recent functional imaging studies of pitch processing claiming a special function of the posterior parts of the right supra-temporal region during the perception of tones [13, 20].

5. Discussion

As shown here and by other authors [19, 18, 7], the CPS is a stable marker of the perception of prosodic boundaries. The ERPs regarding the processing of sentence material comprising phonemic and prosodic information thereby clearly indicate, that the CPS has to be noted as an electrophysiological correlate of the perception of Intonational Phrase boundaries (IPh), but not of hierarchically lower-ordered Phonological Phrase boundaries (PPh). However, naive listeners do not seem to be able to identify the relevant cues to structure acoustic input sufficiently in the absence of phonemic information. This result is contrary to a study of Steinhauer and Friederici [18]. But, in contrast to the second ERP study presented here additional phonemic information was always provided by secondary visual input. Furthermore, the participants in this experiment were not naive listeners, but were instructed and intensively trained to solve the task. By truely naive listeners the delexicalized input can obviously not be chunked into prosodic units due to the absence of this phonemic information in the speech stream.
The reason for this failure could be sought in the violation of the "Sense Unit Condition" [14] for IPh's. Lacking any segmental content, the proposed IPh's in the delexicalized material are no longer structural units, and hence cannot be perceived as such. On the other hand, the presence of the frontal negativities for the delexicalized material in the second ERP study indicates a process trying to compensate for this deviant input. One possibility to interpret this large electrophysiological response is the assumption of an increased processing load on the system.
Concurrently, the reviewed fMRI experiment [10] also observed that whenever delexicalized speech was presented, the regional blood supply to fronto-opercular cortices and fronto-subcortical areas increased (see Figure 5). In accordance with our interpretation regarding the ERPs at frontal electrodes, these data was also interpreted as reflecting an increased effort of the speech processing system to cope with the deviant input [10, 11].
Furthermore, also a hemispheric lateralization of the perception of differing structural cues is supported by these data. As also evident from other functional localization studies [5, 21], the right fronto-opercular cortex is more strongly activated by the detection of pitch variation in the incoming sound stream, whereas the left front-opercular region is related stronger to the extraction of the segmental information.
Whether the observed activations in the frontal brain regions are consistent with the neural generators of the increased negativity in the ERPs, can only be hypothesized here. For clarification of this matter another study would have to be conducted respecting the methodological restrictions of both fMRI and ERP at the same time.
Additional experiments are also necessary to investigate in more detail, how the human brain basically processes the different functions of acoustic vs. prosodic information. For this purpose, special manipulations such as Harmonics-to-Noise filters might be useful.
Delexicalization procedures represent only one step into the direction of complete segregation of prosodic parameters during auditory language processing. Successively, Harmonics-to-Noise filters could manipulate separately the harmonic and spectral properties of a complex speech signal.
The employment of filtered speech material could also lead to a better understanding of the right hemisphere's function in language processing.
A separate manipulation of the harmonic features contained in naturally produced speech then allows to investigate, whether the hemispheric laterality observed so far is due to temporal and/or frequency processing.
More fine-grained, by stepwise application of spectral filtering, the right hemisphere's sensitivity to the perception of different frequency properties in the speech signal (if there is any) could be uncovered (see also Marc Pell's discussion of the role of the right hemisphere in this volume).
It is also conceivable to employ a reverse manipulation, which exclusively extracts pitch information from natural speech. Whispering speech represents a very natural manipulation to receive such a result. Unfortunately, the realization of this last intention is quite problematic due to the loudness of the scanner during data acquisition in fMRI experiments.

6. References

[1] Baum, S.; Pell , M., 1999. The neural basis of prosody: Insights from lesion studies and neuroimaging. *Aphasiology*, 13, 581-608.

[2] Cooper, W.; Paccia-Cooper, J., 1980. *Syntax and Speech.* Cambridge, MA: Harvard UP.

[3] Cutler, A.; Dahan, D.; van Donselaar, W., 1997. Prosody in the comprehension of spoken language. *Language and Speech*, 40, 141-201.

[4] Fodor, J.D., 2002. Psycholinguistics cannot escape prosody. *submitted*.

[5] Gandour, J., 2000. A crosslinguistic PET-study of tone perception. *Journal of Cognitive Neuroscience*, 12, 207-222.

[6] Hagoort, P.; Brown, C., 1999. *The Neurocognition of language.*, New York: Oxford UP.

[7] Hruska, C.; Alter, K.; Steinhauer, K. et al., 2001. Misleading dialogues: Human‘s brain reaction to prosodic information. *Paper presented at the ORAGE-conference*, Aix-en-Provence.

[8] Kutas, M., 1997. Views on how the electrical activity that the brain generates reflects the functions of different language structure. *Psychophysiology*, 34, 383-398.

[9] Leuckefeld, K.; Hahne, A.; Alter, K., 2001. Neuronal Correlates of Processing Intonational Phrase Boundaries in School-aged Children. *Poster presented at the workshop on Prosody in Processing.*, Utrecht.

[10] Meyer, M. et al., 2002. Functional MRI reveals brain regions mediating slow prosodic modulations in spoken sentences. *Human Brain Mapping*, 16, 4, to appear.

[11] Meyer, M.; Alter, K.; Steinhauer, K. et al., 2001. Cerebral substrates of pitch modulations in sentences: Evidence from 3T. *Journal of Cognitive Neuroscience (Supplement)*, 158.

[12] Oldfield, R., 1971. The Assessment analysis of handedness: The Edinburgh Inventory. *Neuropsychologia*, 9, 97-113.

[13] Perry, D. et al., 1999. Localization of cerebral activity during simple singing. *Neuroreport*, 11, 3979-3984.

[14] Selkirk, E., 1986. On derived domains in sentence phonology. *Phonology Yearbook*, 3, 371-405.

[15] Selkirk, E., 1984. *Phonology and Syntax: The Relationship between Sound and Structure,*, Cambridge: MIT Press.

[16] Selkirk, E., 1978. On prosodic structure and its relation to syntactic structure. In *Nordic Prosody II*, Fretheim, T. (ed.). Trondheim: TAPIR.

[17] Sonntag, G.; Portele, T., 1998. PURR- a method for prosody evaluation and investigation. *Journal of Computer Speech and Language*, 12, 437-451.

[18] Steinhauer, K.; Friederici, A., 2001. Prosodic boundaries, comma rules and brain responses: The Closure Positive Shift in the ERPs as a universal marker for prosodic phrasing in listeners and readers. *Journal of Psycholinguistic Reserach*, 30, 267-295.

[19] Steinhauer, K.; Alter, K.; Friederici, A., 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, 2, 191-196.

[20] Tsourio, N., 1997. Functional anatomy of human auditory attention studied with PET. *Neuroimage*, 5, 63-77.

[21] Zatorre, R. et al., 1994. Neural mechanisms underlying melodic perception and memory for pitch. *Journal of Neuroscience*, 14, 1908-1919.