Psychophysiology, 41 (2004), 259–268. Blackwell Publishing Inc. Printed in the USA.
Copyright © 2003 Society for Psychophysiological Research
DOI: 10.1111/j.1469-8986.2003.00142.x

Recognition of affective prosody: Continuous wavelet measures of event-related brain potentials to emotional exclamations

VLADIMIR BOSTANOV AND BORIS KOTCHOUBEY
Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, Tübingen, Germany

Abstract

The affective state of a speaker can be identified from the prosody of his or her speech. Voice quality is the most important prosodic cue for emotion recognition from short verbal utterances and nonverbal exclamations, the latter conveying pure emotion, void of all semantic meaning. We adopted two context violation paradigms—oddball and priming—to study the event-related brain potentials (ERP) reflecting this recognition process. We found a negative wave, the N300, in the ERPs to contextually incongruous exclamations, and interpreted this component as analogous to the well-known N400 response to semantically inappropriate words. The N300 appears to be a real-time psychophysiological measure of spontaneous emotion recognition from vocal cues, which could prove a useful tool for the examination of affective-prosody comprehension. In addition, we developed a new ERP component detection and estimation method that is based on the continuous wavelet transform (CWT), does not rely on visual inspection of the waveforms, and yields larger statistical difference effects than classical methods.

Descriptors: Prosody, Event-related potential, N300, N400, Wavelet, Single trial

"In the beginning was the Scream and the Scream became a Word."[1] This may well be the evolutionary theorist's version of the story about the genesis of language. All birds and mammals have a large repertoire of vocal and bodily signals or displays to express their motives and emotions (e.g., Darwin, 1872; Scherer & Kappas, 1988). Primates show some elements of semantic communication, for instance, uttering different cries of fear in the face of different kinds of danger (Seyfarth, Cheney, & Marler, 1980). Chimpanzees are even capable of learning some elementary propositional language and rudimentary abstract thinking (Premack, 1971). Yet, humans are the only speaking creatures. Even they, however, use nonverbal utterances rather than words to express intense or sudden emotions (Goffman, 1978). Humans emit screams, squeaks, moans, growls, and so forth, termed "sounds of nature" (Naturlaute), "interjections," "emotional vocalizations," "response cries," or "affect bursts," and these sounds, together with facial expressions and bodily gestures, comprise the most natural and ancient language of emotion communication whose expressiveness no words can ever achieve (Scherer, 1985, 1994). These behavioral patterns appear to be quite similar in humans and lower primates, as research indicates that naive human listeners can identify correctly some basic emotions from monkey vocalizations (Leinonen, Linnankoski, Laakso, & Aulanko, 1991; Linnankoski, Laakso, Aulanko, & Leinonen, 1994). Although nonverbal affective exclamations are used relatively rarely in everyday human social communication, spoken language still possesses a variety of nonverbal means to express emotions (Banse & Scherer, 1996). These paralinguistic features make up the affective or emotional prosody of speech.[2]

The ability to identify a speaker's emotions from prosodic cues is a distinct cognitive function that is often impaired in neurological patients with right brain damage (RBD; Ross, Thompson, & Yenkosky, 1997). Neuropsychological research

We thank Simone Lang for recording the emotion names in the priming experiment and Jürgen Mellinger for his valuable recommendations for the acoustical analysis. V.B. gratefully appreciates the inspiring discussions with Eveline Cioflec from the Philosophy Department of the University of Tübingen. This work was supported by the German Research Society (DFG, SFB 550, B3) for B.K., the Institut für Grenzgebiete der Psychologie und Psychohygiene at Freiburg, Germany, the Fortüne Research Program of the Medical School of the University of Tübingen (grant No. 1066-0-0 for B.K., June 2002) and by the Federal Ministry for Science and Technology.

Address reprint requests to: Boris Kotchoubey, Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, Gartenstr. 29, D-72074 Tübingen, Germany. E-mail: boris.kotchoubey@Uni-Tuebingen.DE.

[1] In analogy with "In the beginning was the Word, and the Word was with God, and the Word was God." John 1.1. *The New Oxford Annotated Bible*. (1991). New Revised Standard Version. New York: Oxford University Press.

[2] Affective prosody stands in opposition to linguistic prosody, which serves semantic purposes (e.g., discriminating an assertion from a question).

has shown that certain right hemispheric lesions may even cause complete loss of this ability (i.e., sensory aprosodia) without any significant impairment of other cognitive functions (Ross, 1981). A double dissociation also has been found between comprehension of emotional prosody and identification of affective facial expressions (Adolphs & Tranel, 1999; Anderson & Phelps, 1998). The former function is associated with the right posterior sylvian cortex (Ross, Orvelo, Burgard, & Hansel, 1998), and the latter with the amygdala (Adolphs, Tranel, Damasio, & Damasio, 1994). Emotional-prosody comprehension is a cognitive function of high clinical importance. Besides in RBD, it has been found to be impaired in various neurological and psychiatric disorders that include Parkinson's disease (Breitenstein, Daum, & Ackermann, 1998), Huntington's disease (Speedie, Brake, Folstein, Bowers, & Heilman, 1990), schizophrenia (Ross et al., 2001), alcoholism (Monnot, Nixon, Lovallo, & Ross, 2001), depression (Emerson, Harrison, & Everhart, 1999), attention-deficit hyperactivity disorder (Manassis, Tannock, & Barbosa, 2000), and alexithymia (Lane et al., 1996).

Unfortunately, the term "prosody" is often understood in a narrow sense, referring to the suprasegmental features of speech defined only on a large time scale by pitch and loudness contours and speech and articulation rates (Breitenstein et al., 1998; Crystal, 1969). In this sense, it does not include voice quality, which is defined on a small time scale and is acoustically described by the spectral characteristics (timber) of the vocal sound signal, as well as by rapid, possibly aperiodical changes in frequency that are not perceived as inflections, but rather as specific vibrations or harshness of the voice (Laver, 1980). Clearly, the shorter an emotional utterance (verbal or nonverbal), the fewer prosodic features are present in the narrow sense of the term, and the more important the voice quality is for affect recognition. Thus, it is reasonable to hypothesize that short exclamations convey emotion predominantly through voice quality. Furthermore, Scherer (1986) suggested that voice quality is the key to the differentiation of emotions in affective speech. On the other hand, "sensory aprosodia" denotes impaired comprehension of *all* affective components of language, including emotional voice quality (Ross, 1981). RBD patients were found equally impaired in affect recognition from sentences and single-segment nonverbal utterances ("aaaahhhhhhh") as tested by the Aprosodia Battery (Ross et al., 1997). This finding suggests that they missed all affective prosodic cues, including voice quality, which is probably the main medium of emotion in short unisegmental vocalizations. Phonetics defines voice quality in terms of settings. One definition of a setting is a long-term average configuration of the vocal organs, biasing all speaker's phonation and articulation for a certain time period (Laver, 1980). Thus, settings cause audible variations of those acoustical parameters that constitute the physical description of voice quality. Settings can be controlled either deliberately by the speaker or reflexively by physiological changes (e.g., muscle tension, etc.) or they can be triggered by the affective state of the individual (Laver, 1980; Scherer, 1986, 1994). In the latter case, we hypothesize that the same settings are at work in nonverbal and verbal utterances and that emotional vocalizations and emotional speech share the same or similar voice quality and, possibly, other prosodic features as well.

In summary, there are three important features of nonverbal emotional exclamations. First, they convey the speaker's affective state much better than words. Second, they most probably abstract some key prosodic features, primarily the voice quality, of verbal emotional speech. Third, the underlying physiological mechanisms are similar in all primates.

Experimental Studies

Event-related brain potentials (ERP) have proved extremely useful in studying normal (Kutas & Hillyard, 1980) and impaired (Hagoort, Brown, & Swaab, 1996) semantic speech comprehension, as well as the processing of linguistic prosody (Steinhauer, Alter, & Friederici, 1999). ERP components with latencies ranging from about 200 ms to 500 ms are well suited for testing affect recognition from short, one-syllable verbal or nonverbal utterances because voice quality is present and perceivable from the very onset of a vocalization. Only a few studies, however, have employed ERPs to investigate recognition of emotional voice quality in particular and affective prosody in general. Twist, Squires, Spielholz, and Silverglide (1991) investigated ERPs to emotional prosodic stimuli in an oddball task. The participants listened to one-syllable words spoken with neutral (frequent stimuli) and surprised (rare, target stimuli) intonations. They were instructed to press a button on the occurrence of the rare, emotional stimuli. The P300 ERP component to the targets was shown to exhibit a diminished amplitude and delayed latency in RBD patients compared to left BD patients and healthy controls. In contrast, Erwin et al. (1991) tested emotion recognition in autistic participants and found a surprisingly normal P300 response to the rare targets in an oddball task with happy and angry prosodic stimuli. Because only two different stimuli were presented in the latter experiment, participants may have discriminated them merely by their physical differences and occurrence frequencies (see below). Other researchers have addressed lateralization issues in affective prosody processing. Erhan, Borod, Tenki, and Bruder (1998) investigated ERPs to emotionally spoken nonsense syllables in a dichotic listening task and found larger amplitudes of the N100 and a slow negativity over the left than the right hemisphere, regardless of the ear in which the syllables were presented. Emotion-related differences were found in a very late (1,500–3,000 ms after stimulus presentation) positive wave that was larger following emotionally negative than positive stimuli. Pihan, Altenmuller, and Ackermann (1997), in contrast, found right-lateralized direct-current components of the EEG during listening to sentences with different emotional prosody, as compared with emotionally neutral sentences. This lateralization changed its sign when subjects were required to repeat the sentences using inner speech (Pihan, Altenmuller, Hertrich, & Ackermann 2000).

The main purpose of the present study was to find a reliable, real-time psychophysiological measure of *immediate* and *spontaneous* recognition of affective prosody, specifically emotional voice quality. Using nonverbal vocalizations as stimulus material should also prevent the production of any linguistic prosody on the part of the speaker, which may confound the subsequent acoustical analysis of affective prosody (Leinonen, Hiltunen, Linnankoski, & Laakso, 1997). On the part of the listener, it forestalls any semantic processing that might interfere with emotion recognition. Moreover, as noted earlier, emotional exclamations are a subset of language per se and hence deserve special attention in their own right.

An active oddball task (with an instruction defining the targets) has two important limitations: It primarily tests discrimination rather than recognition, and discrimination can be achieved mainly on the basis of physical differences between two acoustic signals without making much use of their emotional connotation. The poorer performance of the RBD participants reported by Twist et al. (1991) does suggest, however, at least some affect discrimination along with physical discrimination in the active prosodic oddball. For the purposes of the present investigation of emotion recognition, we adopted paradigms based on context violation rather than on stimulus infrequency or novelty. An important feature of meaningful stimuli is that they can build a context that induces some expectations about the next stimulus to occur. For instance, when one hears the sentence stem "He spread the warm bread with…," one expects the usual ending, "butter." When the word "socks" is presented instead, the cognitive processes related to context violation are reflected by the N400 component of the ERP (Kutas & Hillyard, 1980). This effect has been replicated many times with various types of visual and acoustic stimuli, including written and spoken words and sentences (Bentin, Kutas, & Hillyard, 1993; Bentin, McCarthy, & Wood, 1985; Kutas & Hillyard, 1980; McCallum, Farmer, & Pocock, 1984), pictures of objects (McPherson & Holcomb, 1999) and human faces (Jemel, George, Olivares, Fiori, & Renault, 1999), and environmental sounds (Van Petten & Rheinfelder, 1995).

Oddball paradigm. If in a passive oddball paradigm that does not involve a specific task two categories of stimuli are presented instead of two single stimuli, the rare stimuli may elicit an N400 rather than a P300 (Schlaghecken, 1998). The categories of stimuli must be associated with some meaning, so that the frequent stimuli generate a context and provoke some expectations that are then broken by the occurrences of the deviant stimuli. By the same mechanism, an N400 to rare, incongruous, nontarget stimuli can be found during an active oddball task (Bentin, Mouchetant-Rostaing, Giard, Echallier, & Pernier, 1999). Notably, when several categories are presented with equal probabilities, the target category elicits a delayed P300, but no N400 (Kotchoubey & Lang, 2001), because no consistent context can be generated in this condition.

The first experiment of the present study was a passive oddball with emotional vocalizations. The category of the frequent stimuli comprised four exclamations of joy: "Yeah!" "Heey!" "Wow!" and "Oooh!" A single exclamation of woe: "Oooh!" served as the deviant stimulus. All five exclamations had the same occurrence frequency of 20%. The same vowel "o" was purposefully chosen for both the sad exclamation and one of the joyful exclamations; without an active instruction to assign one of the vocalizations as a target, there was not a clearly defined rare stimulus. Thus, the expressed emotion was the only feature distinguishing "Oooh!" (woe) as deviant, although the context violation effect was readily discernable, as listening to the randomized stimulus sequence gave the impression of a man who was expressing his joy, but now and then unexpectedly uttered a sound of despair and deep sorrow.

The experiment still shared some of the shortcomings of a standard oddball paradigm. First, emotional differences were inevitably expressed by some physical parameters that allowed for the possible contribution of "mechanical" discrimination rather than just emotion discrimination. Second, only two emotions were included, which raises the question of whether woe was really spontaneously recognized or if it was merely inferred by discrimination, possibly based more on different arousal levels than on valence (Banse & Scherer, 1996). Third, because joy is a positive emotion and woe is a negative one, the question arises as to whether the ERP to the rare stimulus did not, in fact, reflect motivational evaluation but, instead, recognition processes (Ito, Larsen, Smith, & Cacioppo, 1998).

Priming paradigm. The second experiment was aimed to replicate the results of the first experiment and to address the three open questions formulated above. We adopted a paradigm from a study by Van Petten and Rheinfelder (1995) that demonstrated the conceptual relationships between spoken words and environmental sounds. In that study, an N400 was found to sounds preceded by inconsistent words (e.g., the sound of helicopter rotor [target] preceded by the word "dog" instead of "helicopter" [prime]). In the present study, we presented different spoken emotion names as primes and corresponding emotional vocalizations as targets. The inconsistent combinations were: "joy"-[grief], "pleasure"-[rage], "surprise"-[disappointment], "disappointment"-[surprise], "grief"-[joy], "disgust"-[terror], "rage"-[fright], "fright"-[pleasure], and "terror"-[disgust], with square brackets denoting the expression of emotion. By taking words rather than vocalizations as primes, we forestalled any possible priming by physical features. The inclusion of a number of different emotions was introduced to engender more recognition and less simple discrimination. Finally, by constructing three types of inconsistent pairs: positive-[negative], negative-[positive], and negative-[negative], we attempted to check whether the elicited ERP response is valence-specific.

Continuous Wavelet Transform

The second purpose of the present study was to develop an estimation method based on the continuous wavelet transform (CWT; Ende, Louis, Maass, & Mayer-Kress, 1998; Samar, Bopardikar, Rao, & Swartz, 1999) with the goals of reducing the variance of the obtained ERP measure, developing a procedure suitable for single-trial assessments, and improving ERP component detection from average waveforms. CWT is a mathematical transformation that maps time curves onto smooth, two-dimensional surfaces that are called scalograms.[3] Computationally, the CWT is obtained from the cross-covariance of the ERP curve with a given template function that is described as a wavelet and is systematically varied in width (scale) and position in time. The local extrema of the CWT provide the template's scale and time positions that best match by offering the highest covariance with the ERP curve. Thus, a peak in the curve is represented by a peak in the scalogram, where the two-dimensional position of the scalogram indicates latency and width of the curve. In this sense, CWT is similar to some classical template-matching algorithms used in single-trial ERP analysis (Smulders, Kenemans, & Kok, 1994). An advantage of CWT, however, is the special form of the wavelet template, which allows for optimal *scale separation* and hence better distinction of overlapping ERP components.

Methods

Participants

There were 19 participants (10 men and 9 women; mean age = 26 years) in the oddball experiment and 29 participants (16 men and 13 women; mean age = 23 years) in the priming experiment. All participants were right-handed and their native language was German, as assessed by self-report. The majority of participants

[3] Whereas the term *CWT* can denote both the procedure and its results, the term *scalogram* refers to the result of the transformation.

were students and the remainder were university employees and their relatives. All participants were paid for their participation at a rate of 15DM per hour. None of the participants in the first experiment took part in the second one.

Stimuli

All stimuli were recorded digitally at 22.05 kHz and at a 16-bit sampling rate. All exclamations were uttered by the first author, who is a male, nonnative German speaker (Bulgarian) and all of the emotion names were spoken by a female native German speaker. Neither of the speakers was a professional actor. Stimulus duration varied from 750 ms to 870 ms in the oddball paradigm and from 630 ms to 980 ms in the priming experiment. In the oddball paradigm, the duration of "Oooh!" (woe) and "Oooh!" (joy) was approximately the same duration at about 840 ms. In the priming paradigm, the longest emotion name, "disappointment" (German: "Enttäuschung") lasted approximately 790 ms.

Procedure

In the oddball experiment, the five exclamations were presented 60 times each in a randomized sequence of 300 trials, at a rate of 1 stimulus per 1.1 s. The word-exclamation pairs in the priming experiment were presented in a randomized sequence of 108 pairs; the nine exclamations were repeated 12 times—6 times preceded by the correct emotion name and 6 times preceded by a wrong name. The stimulus onset asynchrony between a word and an exclamation was 1 s and the presentation rate was 1 pair per 3 s. Digitized EEG [time resolution: 2 ms/step (500 Hz), voltage resolution: 0.1678 μV/step] was continuously recorded from nine scalp positions according to the 10-20 system: Fz, Cz, Pz, F3, F4, C3, C4, P3, P4. During the oddball experiment, all electrodes were referenced to linked mastoids; during the priming experiment, the nose was used as the reference and subsequent off-line rereferencing to mastoids did not yield any observable change in results. Vertical and horizontal eye movements were recorded from FP2 and a site below the right eye and from F7 and F8, respectively.

In both experiments, the only instruction given to the participants was to listen attentively. At the end of each experiment, participants were asked to provide detailed verbal reports of what they had heard.

Data Analysis

The acoustic features of the oddball stimuli were analyzed with the short-time Fourier transform (STFT). The EEG was filtered on-line (bandpass: 0.1 Hz–70 Hz, notch: 50 Hz). EEG epochs were created off-line. Eye-blink and eye-movement artifacts, both vertical and horizontal, were corrected off-line by a computerized procedure (Gratton, Coles, & Donchin, 1983; Miller, Gratton, & Yee, 1988). Trials with voltage exceeding 90 μV in any channel were considered to be contaminated by artifact and were excluded from further analysis. Averages for each participant as well as grand averages over all participants were calculated for each experimental condition using a 100-ms prestimulus baseline. CWTs were computed for each electrode, for each participant's average waveform as well as for each single trial, using the "Mexican Hat" wavelet (Ende et al., 1998). Three different ERP measures were obtained from participant averages and single trials: area, fixed wavelet, and matched wavelet.

An *area* was computed as the average voltage in a time window determined by visual inspection of the grand average waveforms. The area measure can be considered as a form of template matching. It is computationally equivalent to the cross-covariance of each participant average and each single trial with a fixed rectangular template[4] (see Figure 1) that is fitted visually to an ERP component (also detected visually) in the grand-average waveforms, without any additional fitting of the template's or window's position and width to individual participant averages or single trials.

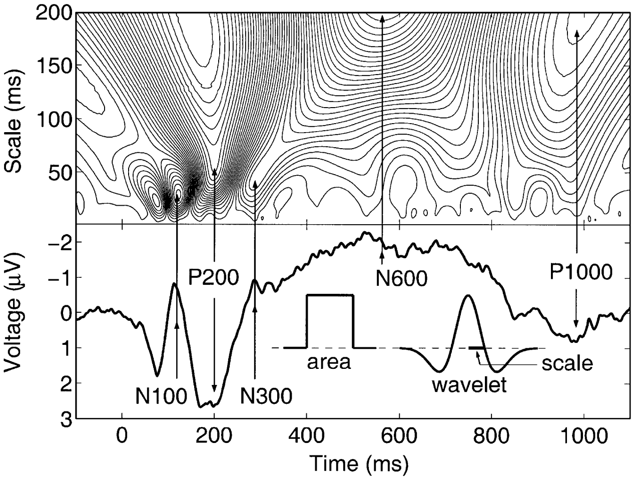

Figure 1. The rectangular area template, the wavelet template, the total grand average of both experimental conditions in the oddball experiment (lower plot), and its CWT scalogram (upper plot). Each point of the scalogram represents the cross-covariance of the total grand average with a wavelet of the corresponding time and scale. Extrema in the scalogram correspond to peaks in the waveform. Extremum scale gives a measure of peak width (ERP component duration). The scale of the Mexican Hat wavelet (lower plot) is defined as the half-width between its zeros (thick horizontal line). Note that the area template has no minima under the zero line and hence cannot separate larger scales.

A wavelet measure was defined as a CWT value (i.e., the cross-covariance of the waveform with a wavelet) at a certain time and scale, according to the following procedure. First, a *total grand average* was calculated for each experiment. The total grand average included *all trials* obtained from all participants during *both* experimental conditions. Next, the CWT scalogram of the total grand average at Cz was computed, and ERP components were identified as local extrema in this scalogram. As shown in Figure 1, the position of the N300 minimum was approximately 290 ms in the time domain and 45 ms in the scale domain (which corresponds to 22.2 Hz in the frequency domain). The N300 *fixed-wavelet* measure was defined as the CWT value at this same position for all participant averages, all single trials, and all channels. The N300 *matched-wavelet* measure was defined as the minimum CWT value in the rectangular window 270 ms < time < 310 ms, 35 ms < scale < 55 ms in each participant-average and each single-trial scalogram. Thereby, the matched wavelet was partly adjusted to compensate for inter- and intraparticipant latency variations. Thus the fixed wavelet was fitted only once to the total grand average and then the cross-

[4] Strictly speaking, this is only true under the assumption that the mean of an EEG curve is zero, which may be wrong for single (baseline-corrected) sweeps; however, this does not lead to loss of generality of our reasoning. In the case of a wavelet template, this assumption is unnecessary, because the mean of a wavelet is zero by definition.

covariance with each participant average and each single trial was taken as a measure of the N300 effect, whereas the matched wavelet was additionally fitted to each participant-average and each single-trial waveform.

All ERP measures were tested for statistical significance by three-way analyses of variance (ANOVA) with factors: condition (2) × frontal/central/parietal (FCP; 3) × left/mid/right (LMR; 3). The experimental conditions of interest were "Oooh!" (joy)/ "Oooh!" (woe) during the oddball paradigm and consistent/ inconsistent during the priming paradigm. FCP and LMR were taken as within-subject or within-trial factors, whereas the experimental condition was taken as a within-subjects factor for the participant-average ERP measures and as a between-trials factor for the single-trial ERP measures. Left/right and frontal/ parietal asymmetries were studied by linear contrasts; predominance at the vertex Cz was studied by quadratic contrasts. (With threefold topography factors, a main effect or an interaction with the condition factor may reflect either an asymmetry or a central deviation, or both; contrasts specify which of the three is true.)

The wavelets were computed with MATLAB 6.0, and the ANOVAs with SPSS 11.0.

Results

For the sake of brevity and clarity, we report only significant ANOVA probability values. Similarly, results obtained with the fixed-wavelet measure are not reported and the "matched wavelet" that produced generally better results is subsequently referred to as "wavelet" only. Although not all four ERP measures (area and wavelet, both single-trial and participant-average) are needed to characterize the components, their comparison is useful for testing the new total-average-CWT assessment method.

Oddball Paradigm

Verbal reports. All participants correctly identified the emotions expressed by the presented exclamations. "Yeeh!" "Heey!" "Wowh!" and "Oooh!" (joy) were described as "cheerful," "joyful," "merry," "expressing pleasure," "positive surprise," or "admiration." "Oooh!" (woe) was described as "mourning," "suffering," expressing "grief," "desperation," or "pain."

Acoustical analysis. Figure 2 shows that emotional voice quality, and more precisely emotional timber, was the only prosodic feature that distinguished "Oooh!" (woe) from the joyful exclamations.

ERP measures. Grand-average waveforms are shown in Figure 3. N300 amplitude was significantly larger for "Oooh!" (woe) than for "Oooh!" (joy) as revealed by a significant main effect for condition for the participant-average measures: area, $p = .03$, and wavelet, $p = .001$, as well as with the single-trial wavelet, $p < .001$, but not with the single-trial area. The component was predominant at Cz (see Figure 4) as significant quadratic trends were obtained for FCP, LMR, and FCP × LMR for both wavelet measures: participant-average, $p = .005$, and single-trial, $p < .001$, but not for the area measures. N600 amplitude was significantly larger for "Oooh!" (woe) than for "Oooh!" (joy) at frontal sites for the Condition × FCP interaction for both participant-average measures: area, $p = .03$, and wavelet, $p < .001$, as well as for the single-trial wavelet, $p < .001$, but not for the single-trial area. N600 was predominant at midline sites as revealed by a significant quadratic effect for LMR across all of the measures: participant-average: area, $p = .02$, and wavelet, $p = .001$, and single-trial: area, $p = .008$, and wavelet, $p < .001$.

P1000 amplitude was significantly larger for "Oooh!" (woe) than for "Oooh!" (joy) at frontal sites, as significant main effects for condition were obtained with both wavelet measures: participant-average, $p = .03$, and single-trial, $p = .01$, but not the area measures; significant Condition × FCP interaction effects also were observed for both participant-average measures: area, $p = .01$, and wavelet, $p < .001$, as well as the single-trial wavelet, $p < .001$, but not the single-trial area.

Priming Paradigm

Verbal reports. All participants in the priming experiment correctly identified consistent and inconsistent word exclamation pairs and reported that each of the nine emotional words was

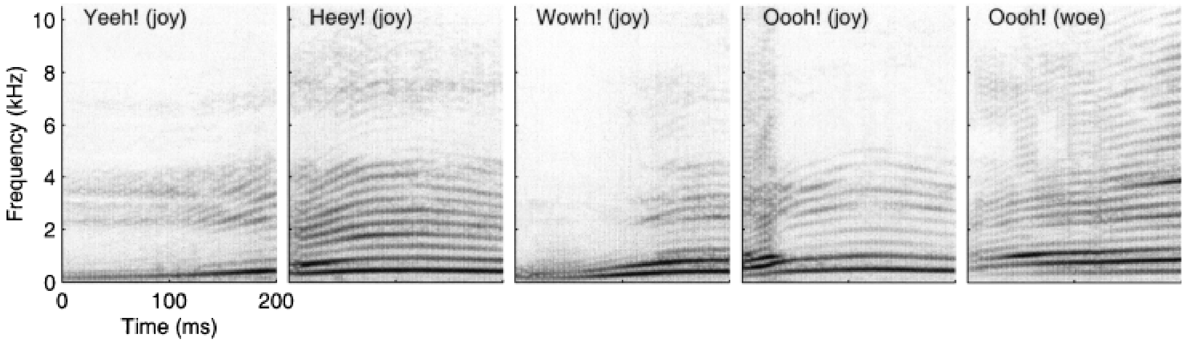

Figure 2. STFT spectrograms of the first 200 ms of each of the exclamations in the oddball experiment. Darker shades indicate higher intensities. The first dark stripe from the bottom represents the fundamental frequency f0, the second represents the first harmonic f1, and so forth. Two principal spectral parameters discriminating between the emotions are readily discernable: First, the energy distribution between f0 and f1 is in favor of f0 for joy and clearly in favor of f1 for woe, and second, the proportion of energy contained in frequencies above f11 (about 4.5 kHz) is also fairly larger for woe than for joy. Other acoustic parameters—f0-mean (pitch), f0-contour (intonation), amplitude envelope (accentuation)—vary among all exclamations and have approximately mean values for "Oooh!" (woe).

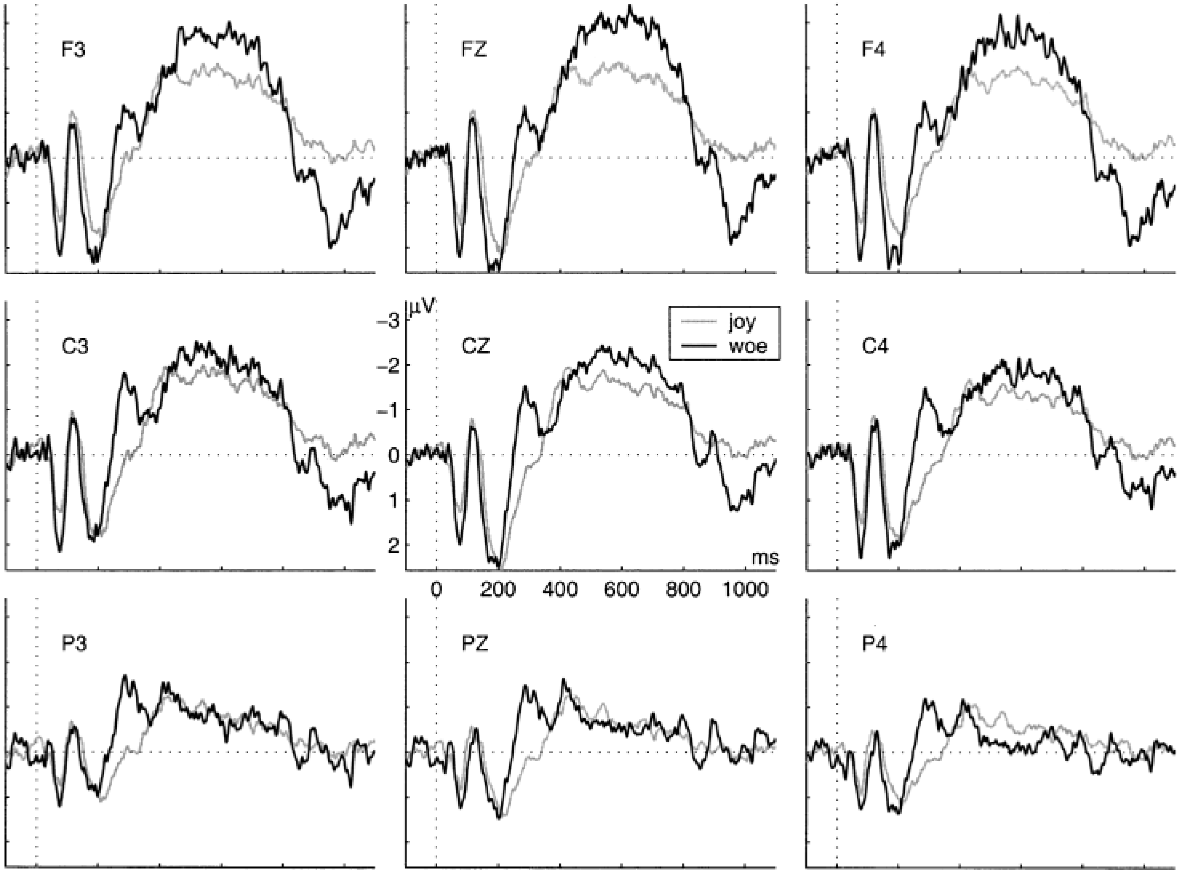

Figure 3. Grand average waveforms for the four joyful exclamations: "Yeeh!" "Heey!" "Wowh!" "Oooh!" (all four averaged together), and for the doleful one: "Oooh!" in the oddball experiment. Besides the P50/N100/P200 complex, three further prominent ERP components can be distinguished: N300—an early negative wave with peak latency ≅ 300 ms, duration ≅ 100 ms, peak-to-peak amplitude ≅ 1.5 μV, and broad scalp distribution; N600—a later, slower negative wave with latency ≅ 600 ms, duration ≅ 400 ms, and centro-frontal distribution; P1000—a late positive wave with latency ≅ 1,000 ms, duration ≅ 150 ms, and centro-frontal distribution.

followed sometimes by the corresponding vocalization and sometimes by an inappropriate one(s).

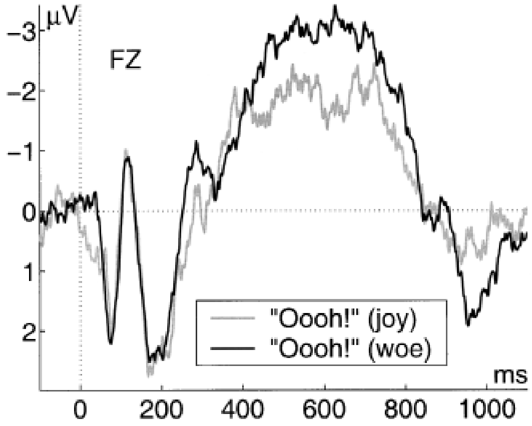

Figure 4. Grand average waveforms at Fz for "Oooh!" (woe) and "Oooh!" (joy) in the oddball experiment. The ERP to "Oooh!" (joy) is almost the same as the average across all four joyful exclamations (see Figure 3).

ERP measures. Grand-average waveforms are shown in Figure 5. N300 amplitude was significantly larger for the inconsistent vocalizations as revealed by a significant main effect of condition for both wavelet measures: participant-average, $p = .02$, and single-trial, $p = .04$; the effect was marginally significant for the participant-average area, $p = .05$, and not significant for the single-trial area. The component was predominant at the vertex Cz, as significant quadratic contrasts were observed for FCP, LMR, and FCP × LMR for both wavelet measures: participant-average, $p < .05$, and single-trial, $p < .001$, but not for the area measures.

P500 amplitude was significantly larger for the inconsistent vocalizations with a significant main effect of condition for both wavelet measures: participant-average, $p = .04$, and single-trial, $p = .006$, but not for the area measures. The component was predominant at parietal sites, as the linear FCP contrast was significant for both wavelet measures: participant-average, $p < .001$, and single-trial, $p < .001$, but not for the area measures. Notwithstanding visual impression, the linear-contrast effect of FCP × condition was not significant.

The significance of the N300 effect was also tested separately for each of the three kinds of word-exclamation pairs: positive-[negative] ("joy"-[grief], "pleasure"-[rage], "surprise"-[disappointment]), negative-[positive] ("grief"-[joy], "fright"-[pleasure],

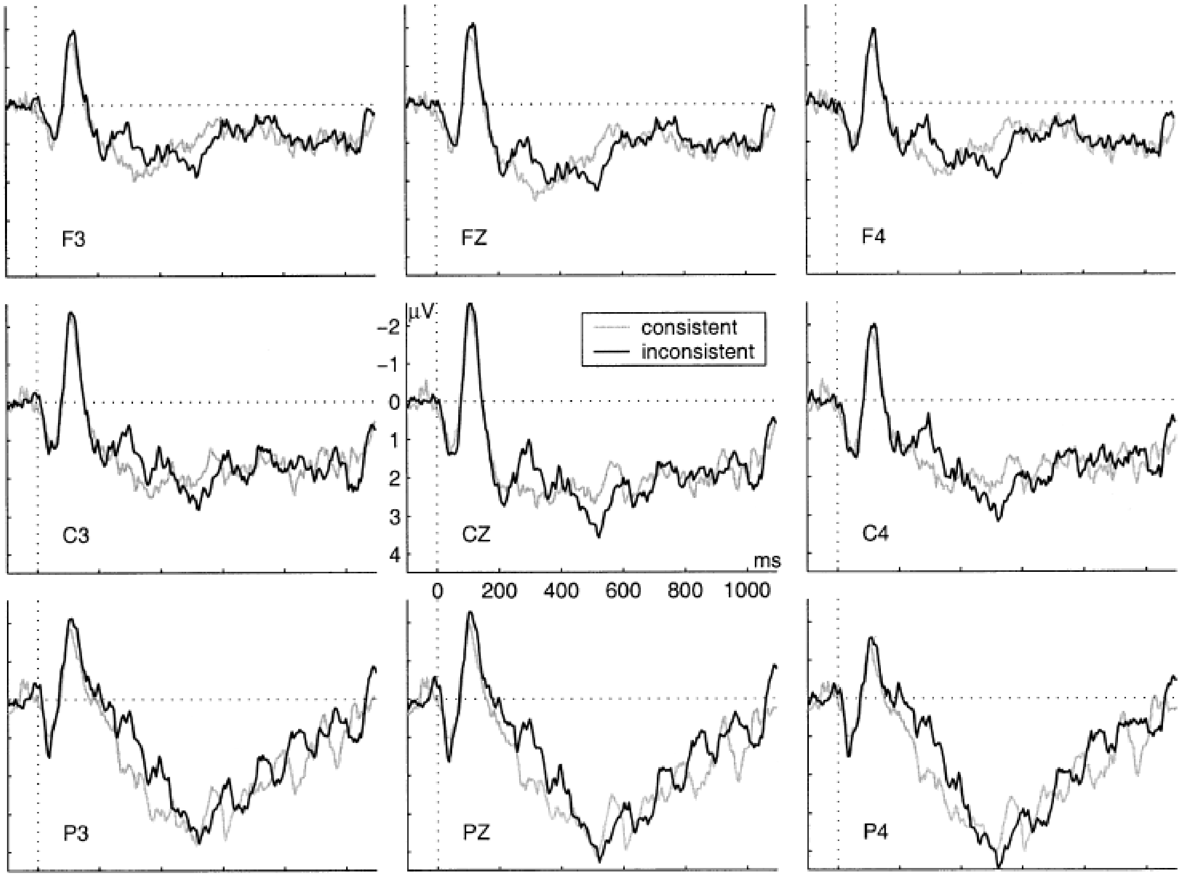

Figure 5. Grand average waveforms for the consistent and inconsistent exclamations in the priming experiment. Besides the P50 and the N100, two further ERP components can be seen: an N300 very similar to the one elicited by the sad exclamation in the oddball, and a P500—a positive wave with latency ≅ 500 ms and duration ≅ 150 ms.

"disappointment"-[surprise]), and negative-[negative] ("disgust"-[terror], "rage"-[fright], "terror"-[disgust]). Both wavelet measures were significant for the negative-[positive] pairs, $p < .05$, and marginally significant for the positive-[negative] and negative-[negative] pairs, $.05 < p < .07$; the area measures were not significant.

Discussion

Recognition of Affective Prosody

The verbal reports confirmed that participants identified all emotions correctly, and results of the acoustical analysis support the hypothesis that the nonverbal vocalizations utilized in the current study conveyed emotion primarily by voice quality. An N300 was elicited both by the deviant emotional exclamations in the passive oddball paradigm and by the inconsistent affect vocalizations in the priming experiment. The N300, therefore, appears to be an early N400-like wave that reflects cognitive processes related to context violations and broken expectations during emotion recognition of the prosodic features of nonverbal stimuli. The occurrence of an N300 to both positive and negative inconsistent exclamations suggests that this component is not valence specific. The short latency and duration of the N300 indicate that affect recognition appears to be rapid and based primarily on voice quality within the first 100–150 ms of exclamations. This result is consistent with an earlier finding of emotion recognition from extremely short voice samples with durations of less than 100 ms (Pollack, Rubenstein, & Horowitz, 1960). The finding of an N400-like response to meaningful infrequent stimuli during a passive oddball paradigm is not surprising, as the frequent joyful exclamations appear to have served as a context background for the incongruous, rare, doleful exclamation.

The design of the present priming experiment most closely parallels an N400 study of environmental sounds by Van Petten and Rheinfelder (1995). They observed an N400 with a similar onset latency of 200 ms, but a much longer duration of 700 ms as compared with the 100-ms duration of the N300 obtained in the present study. This may be due to considerable latency jitter of the negative wave elicited by incongruous sounds. Whereas emotional exclamations are apparently recognized very rapidly by their onset voice quality, other environmental sounds may be impossible to identify from their onset acoustics. This was likely to have been the case for at least one of the sets of sounds described by Van Petten and Rheinfelder (1995): "horse hooves striking pavement." In this example, at least two consecutive hoofbeats need to be processed to classify the stimuli as congruous or incongruous and perhaps even more are necessary for correct identification, particularly given that this sound is not commonplace in our environment. Such hard-to-recognize sounds are likely to elicit an ERP with a much later onset and peak latency than easy-to-recognize sounds, such as affective vocalizations. Because a large number of sounds were used in the study by Van Petten and Rheinfelder, it is quite possible that the resulting ERP component in the average waveform reflected early onset latencies determined by stimuli that were easy to recognize as well as longer durations associated with stimuli that were more difficult to recognize. The resulting ERP peak latency would be determined by the most frequent stimulus recognition latency. Furthermore, the variance in the recognition point across stimuli is not the only possible source of latency jitter. It is also possible that there was considerable variance in identification time between participants, as it is unlikely that all of the sounds were equally familiar to all participants. The opposite may have been true for the emotional exclamations presented in the current study, as the short duration of the N300 suggests that all affective stimuli were identified equally rapidly by all participants.

Studies investigating interstimulus (Woodward, Owens, & Thompson, 1990) and intersubject latency variability (Moreno, Federmeier, & Kutas, 2002) of N400 to semantic violations found that this latency is rarely less than 300 ms even in participants of the highest English proficiency. Comparing these and other semantic priming results to the outcome of our experiments, we can infer that nonverbal affective vocalizations are recognized approximately 100–150 ms earlier. If we assume that the recognition of emotional exclamations is also representative for the recognition of affective prosody of verbal material, we can conclude that emotion may be grasped faster than meaning. This conjecture (which can be tested in an experiment using emotionally spoken words instead of vocalizations) may imply that affect recognition could potentially bias or override further semantic prosody, especially when contradictory messages are carried by the semantics, on the one hand, and the voice quality, on the other hand.

Although this view is similar to the theory of fast emotional responses based on rapid stimulus processing in the amygdala, preceding the slow recognition by the cortex (LeDoux, 1993), it is not identical. As noted in the introduction, a considerable body of evidence indicates that emotional prosody is processed in the cortex rather than in the amygdala. The present hypothesis concerns emotion perception rather than emotional responses. Clearly, the former does not necessarily imply the latter, as an individual who has adequately recognized emotional prosody in another individual can nevertheless remain calm and impassive.

An open question remains as to what extent the affect recognition process was facilitated by learning during the experiments. One possible scenario is that during the earlier (e.g., 10–20) trials, participants identified emotions by the prosodic features manifested in whole vocalizations, and then quickly learned to discriminate familiar stimuli on the basis of their onset. This problem stems from the great number of repetitions of the same few different exclamations. A further question is whether emotion recognition or discrimination from the first 100–150 ms of the vocalizations was based solely on voice quality, or if intonation and accentuation patterns could have been discernable and utilized by participants even though the patterns were not detected by the spectrogram. Apart from a more elaborated acoustical analysis, there are several potential approaches to these problems. One possibility is to use a variety of exclamations generated by many different voices, thereby avoiding repetitions. Another solution is to use digitally manipulated sound signals derived from the same neutral vocalization (Pihan et al., 2000), and a third option is to use vocalizations or emotionally spoken words truncated to 100 ms, for instance, as the stimulus material (Pollack et al., 1960).

Another important question is to what extent nonverbal vocalizations and emotional speech share the same prosodic features. Extensive acoustic analyses are required to clarify this issue. The other ERP components that were found in the current experiments also demand further investigation to clarify their relation to emotion recognition.

The Total-Average-CWT Method

A new CWT-based method was developed in an effort to improve detection and assessment of ERP components from single trials as well as from participant averages. Essentially, it is a type of template matching, where the template is a wavelet, and it is fitted to the signal not only in time (component latency) but also in frequency (scale, peak width). In all assessments, the wavelet measure yielded larger ANOVA effects than the traditional area measure. This effect was particularly pronounced in the single-trial assessments. This result is not surprising, considering that traditional area measures do not separate scales, and hence peaks of different widths, especially slow waves and DC components, are taken into account in a single component assessment, thus increasing variance and reducing statistical effect size and statistical power. This is best understood if the area measure is seen as template matching with a fixed rectangular template, which takes only positive values, whereas the wavelet template with its two deep minima extracts only peaks of nearly the same scale (Figure 1).

Furthermore, the CWT method has three important advantages over its ancestors, the classical template-matching, single-trial algorithms (Smulders et al., 1994). First, the template is fitted to the signal not only in time, but also in scale. Second, the template is a wavelet, which guarantees better scale separation than nonwavelet templates, and hence better selective sensitivity of the wavelet measures to particular ERP components as well as better separation of overlapping components (Samar et al., 1999). Third, the scale and time position of the template are determined from the total average of all trials (or all participant averages) entering the subsequent statistical analyses. The detection of ERP peaks from the total-average scalogram is blind to the separation of the sample into categories according to the experimental conditions. This makes it statistically more appropriate for testing difference hypotheses than traditional methods that use the difference between two waveforms derived from different conditions to identify the relevant components. The post hoc visual selection of time windows from this difference may bias subsequent statistical testing due to accumulation of chance. The total-average scalogram allows unbiased ERP analysis without visual inspection of the curves.

A possible shortcoming of this method arises in the relatively rare instance when separate experimental conditions elicit an ERP component of the same latency, width, and amplitude, but with different polarities, as such waves will cancel each other in the total average. This problem has been resolved in a more recent version of the CWT method, the t-CWT, developed by our group (Bostanov, in press).

REFERENCES

Adolphs, R., & Tranel, D. (1999). Intact recognition of emotional prosody following amygdala damage. *Neuropsychologia*, *37*, 1285–1292.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, *372*, 669–672.

Anderson, A. K., & Phelps, E. A. (1998). Intact recognition of vocal expressions of fear following bilateral lesions of the human amygdala. *NeuroReport*, *9*, 3607–3613.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*, 614–636.

Bentin, S., Kutas, M., & Hillyard, S. A. (1993). Electrophysiological evidence for task effects on semantic priming in auditory word processing. *Psychophysiology*, *30*, 161–169.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*, 343–355.

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: Time course and scalp distribution. *Journal of Cognitive Neuroscience*, *11*, 235–260.

Bostanov, V. (in press). BCI competition 2003—Data sets Ib and IIb: Feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. *IEEE Transactions of Biological Engineering*.

Breitenstein, C., Daum, I., & Ackermann, H. (1998). Emotional processing following cortical and subcortical brain damage: Contribution of the fronto-striatal circuitry. *Behavioral Neurology*, *11*, 29–42.

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge, UK: Cambridge University Press.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: Murray. (Reprinted 1998, London: HarperCollins).

Emerson, C. S., Harrison, D. W., & Everhart, D. E. (1999). Investigation of receptive affective prosodic ability in school-aged boys with and without depression. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*, *12*, 102–109.

Ende, M., Louis, A. K., Maass, P., & Mayer-Kress, G. (1998). EEG signal analysis by continuous wavelet transform techniques. In H. Kantz, J. Kurths, & G. Mayer-Kress (Eds.), *Nonlinear analysis of physiological data* (pp. 213–219). Berlin, Heidelberg: Springer.

Erhan, H., Borod, J. C., Tenke, C. E., & Bruder, G. E. (1998). Identification of emotion in a dichotic listening task: Event-related brain potential and behavioral findings. *Brain and Cognition*, *37*, 286–307.

Erwin, R. J., Van Lancker, D., Guthrie, D., Schwafel, J., Tanguay, P., & Buchwald, J. S. (1991). P3 responses to prosodic stimuli in adult autistic subjects. *Electroencephalography and Clinical Neurophysiology*, *80*, 561–571.

Goffman, E. (1978). Response cries. *Language*, *54*, 787–815.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for offline removal of ocular artifacts. *Electroencephalography and Clinical Neurophysiology*, *55*, 468–484.

Hagoort, P., Brown, C. M., & Swaab, T. Y. (1996). Lexical-semantic event-related potential effects in patients with left hemisphere lesions and aphasia, and patients with right hemisphere lesions without aphasia. *Brain*, *119*, 627–649.

Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personal and Social Psychology*, *75*, 887–900.

Jemel, B., George, N., Olivares, E., Fiori, N., & Renault, B. (1999). Event-related potentials to structural familiar face incongruity processing. *Psychophysiology*, *36*, 437–452.

Kotchoubey, B., & Lang, S. (2001). Event-related potentials in a semantic auditory oddball task in humans. *Neuroscience Letters*, *310*, 93–96.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205.

Lane, R. D., Sechrest, B., Riedel, R., Weldon, V., Kaszniak, A., & Schwartz, G. (1996). Impaired verbal and nonverbal emotion recognition in alexithymia. *Psychosomatic Medicine*, *58*, 203–210.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge, UK: Cambridge University Press.

LeDoux, J. E. (1993). Emotional memory systems in the brain. *Behavioural Brain Research*, *58*, 69–79.

Leinonen, L., Hiltunen, T., Linnankoski, I., & Laakso, M.-L. (1997). Expression of emotional-motivational connotations with a one-word utterance. *Journal of the Acoustic Society of America*, *102*, 1853–1863.

Leinonen, L., Linnankoski, I., Laakso, M.-L., & Aulanko, R. (1991). Vocal communication between species: Man and macaque. *Language & Communication*, *11*, 241–262.

Linnankoski, I., Laakso, M.-L., Aulanko, R., & Leinonen, L. (1994). Recognition of emotions in macaque vocalizations by children and adults. *Language & Communication*, *14*, 183–192.

Manassis, K., Tannock, R., & Barbosa, J. (2000). Dichotic listening and response inhibition in children with comorbid anxiety disorders and ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*, 1152–1159.

McCallum, W. C., Farmer, S. F., & Pocock, P. V. (1984). The effects of physical and semantic incongruities on auditory event-related potentials. *Electroencephalography and Clinical Neurophysiology*, *59*, 477–488.

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology*, *36*, 53–65.

Miller, G. A., Gratton, G., & Yee, C. M. (1988). Generalized implementation of an eye movement correction procedure. *Psychophysiology*, *25*, 241–243.

Monnot, M., Nixon, S., Lovallo, W., & Ross, E. (2001). Altered emotional perception in alcoholics: Deficits in affective prosody comprehension. *Alcohol, Clinical and Experimental Research*, *25*, 362–369.

Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language*, *80*, 188–207.

Pihan, H., Altenmuller, E., & Ackermann, H. (1997). The cortical processing of perceived emotion: A DC-potential study on affective speech prosody. *NeuroReport*, *8*, 623–627.

Pihan, H., Altenmuller, E., Hertrich, I., & Ackermann, H. (2000). Cortical activation patterns of affective speech processing depend on concurrent demands on the subvocal rehearsal system. A DC-potential study. *Brain*, *123*, 2338–2349.

Pollack, I., Rubenstein, H., & Horowitz, A. (1960). Communication of verbal modes of expression. *Language and Speech*, *3*, 121–130.

Premack, D. (1971). Language in chimpanzee? *Science*, *172*, 808–822.

Ross, E. D. (1981). The aprosodias. Functional-anatomic organization of the affective components of language in the right hemisphere. *Archives of Neurology*, *38*, 561–569.

Ross, E. D., Orbelo, D. M., Burgard, M., & Hansel, S. (1998). Functional-anatomic correlates of aprosodic deficits in patients with right brain damage. *Neurology*, *50*(suppl. 4), A363.

Ross, E. D., Orbelo, D. M., Cartwright, J., Hansel, S., Burgard, M., Testa, J. A., & Buck, R. (2001). Affective-prosodic deficits in schizophrenia: Profiles of patients with brain damage and comparison with relation to schizophrenic symptoms. *Journal of Neurology, Neurosurgery and Psychiatry*, *70*, 597–604.

Ross, E. D., Thompson, R. D., & Yenkosky, J. (1997). Lateralization of affective prosody in brain and the callosal integration of hemispheric language functions. *Brain and Language*, *56*, 27–54.

Samar, V. J., Bopardikar, A., Rao, R., & Swartz, K. (1999). Wavelet analysis of neuroelectric waveforms: A conceptual tutorial. *Brain and Language*, *66*, 7–60.

Scherer, K. R. (1985). Vocal affect signaling: A comparative approach. In J. Rosenblatt, C. Beer, M.-C. Busnel, & P. Slater (Eds.), *Advances in the study of behavior* (Vol. 15, pp. 189–244). New York: Academic Press.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychology Bulletin*, *99*, 143–165.

Scherer, K. R. (1994). Affect bursts. In van Goozen, S. H. M., van de Poll, N. E., & J. A. Sergeant (eds.), *Emotions: Essays on emotion theory* (pp. 161–193). Hillsdale, NJ: Lawrence Erlbaum Associates.

Scherer, K. R., & Kappas, A. (1988). Primate vocal expression of affective states. In D. Todt, P. Goedeking, E. Newman, & D. Symmes (Eds.), *Primate vocal communication* (pp. 171–194)). Berlin, Heidelberg, New York: Springer.

Schlaghecken, F. (1998). On processing BEASTS and BIRDS: An event-related potential study on the representation of taxonomic structure. *Brain and Language*, *64*, 53–82.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science*, *210*, 801–803.

Smulders, F. T., Kenemans, J. L., & Kok, A. (1994). A comparison of different methods for estimating single-trial P300 latencies. *Electroencephalography and Clinical Neurophysiology*, *92*, 107–114.

Speedie, L. J., Brake, N., Folstein, S. E., Bowers, D., & Heilman, K. M. (1990). Comprehension of prosody in Huntington's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, *53*, 607–610.

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, *2*, 191–196.

Twist, D. J., Squires, N. K., Spielholz, N. I., & Silverglide, R. (1991). Event-related potentials in disorders of prosodic and semantic linguistic processing 1. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*, *4*, 281–304.

Van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, *33*, 485–508.

Woodward, S. H., Owens, J., & Thompson, L. W. (1990). Word-to-word variation in ERP component latencies: Spoken words. *Brain and Language*, *38*, 488–503.

(Received February 13, 2002; Accepted July 25, 2003)