Rapid processing of emotional and voice information as evidenced by ERPs

Silke Paulmann[1], *Patricia Schmidt*[2], *Marc Pell*[1], *Sonja A. Kotz*[2]

[1] McGill University, School of Communication Sciences & Disorders, Montreal, Canada
[2] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
Silke.Paulmann@mail.mcgill.ca

Abstract

Next to linguistic content, the human voice carries speaker identity information (e.g. female/male, young/old) and can also carry emotional information. Although various studies have started to specify the brain regions that underlie the different functions of human voice processing, few studies have aimed to specify the time course underlying these processes. By means of event-related potentials (ERPs) we aimed to determine the time-course of neural responses to emotional speech, speaker identification, and their interplay. While engaged in an implicit voice processing task (probe verification) participants listened to emotional sentences spoken by two female and two male speakers of two different ages (young and middle-aged). For all four speakers rapid emotional decoding was observed as emotional sentences could be differentiated from neutral sentences already within 200 ms after sentence onset (P200). However, results also imply that individual capacity to encode emotional expressions may have an influence on this early emotion detection as the P200 differentiation pattern (neutral vs. emotion) differed for each individual speaker.

1. Introduction

Whether we like it or not, our voice reveals information about gender, age, social and geographical background, how we feel, and what we really mean. Importantly, the voice can carry all these information types at the same time, thereby making it a powerful instrument that plays a critical role in human communication. These information types are expressed by variations in acoustic parameters, such as tempo, mean amplitude or intensity, and mean fundamental frequency (f0) or pitch. Some authors call the human voice an 'auditory face', meaning that special physical feature combinations are "related to the unique configuration of the human vocal apparatus" [1]. Accordingly, the same authors have suggested a model of voice perception that is comparable to the functional organization of face perception. In this model voice perception is organized hierarchically. A general low-level analysis of auditory information that is mediated by primary auditory cortex regions and subcortical nuclei, is followed by a structural analysis which is supposedly mediated by bilateral regions of the middle superior temporal sulcus (STS). Last, processing of vocal information may then be subdivided into three functionally different processes, involving vocal speech analysis, vocal affect analysis, and vocal identity analysis, with all three processes following different neural pathways [1]. Importantly, the model assumes that these pathways are not independent but may interact at different points in time. In the current paper, we will focus on two of the above listed processes, that is vocal affect and vocal identity analysis, and we will try to specify when these processes first interact.

While many recent neuroimaging studies have specified the brain regions that underlie different functions of human voice processing (for reviews see [1, 2, 3, 4]), the time course of neural responses to vocal affect and vocal identity analysis is less explored. As for the processing of the latter, behavioral studies suggest that voice identity is already processed before phonological encoding occurs [5]. However, ERPs are a more useful tool when investigating temporal aspects of speech processing. The current experiment will test if emotional expressions spoken by female and male speakers will lead to different ERP responses. If so, will it be possible to specify at which point in time this differentiation first occurs? There is evidence from Mismatch Negativity (MMN) studies investigating voice identity irrespective of emotions. For instance, in an oddball experiment, Titova and Näätänen (2001) report an early ERP response for a change in voice identity, suggesting preattentive processing of voice identity features [6]. A similar pre-attentively elicited differentiation effect was reported in an MEG study in which the relationship of voice and linguistic information processing was investigated [7]. Last, our own work supports early speaker gender differentiation as reflected in the P200. We find the amplitude of the P200 varying as a function of speaker voice (male/female). However, interpretation of this effect is limited as we only tested two speaker voices so far [8].

In the same study we report that vocal emotional sentences can be differentiated from neutral sentences as early as 200 ms after sentence onset also reflected in the P200 component [8]. By testing neutral and emotional sentences spoken by a female and a male voice, we explored whether early emotional differentiation varies as a function of speaker voice. While the data suggest that individual capacity to encode vocal emotion may influence vocal emotional processing, they also clearly show that emotional vocalizations can be distinguished from neutral vocalizations very rapidly independent of speaker voice. As mentioned above one limitation of the study was the use of two speaker voices only. To further explore the time course of neural responses to emotional speech, speaker identity, and their interplay we presented stimuli recorded from four different speakers (2 males/2 females) of two different age groups (young, middle-aged) in the current experiment. In particular, we investigated the influence of voice identity on emotional perception to shed more light on the issue of gender and age voice specific emotional processing. For instance, behavioral evidence suggests that emotional *decoding* declines with increasing age [9]. The question arises whether emotional *encoding* also varies as a function of age, that is, is differentiation of vocal emotional and neutral sentences better when expressed by young rather than middle-aged voices? Or is such a differentiation more con-

The authors thank Kerstin Flake for her help with graphical illustrations. This work was supported by the German Research Foundation (DFG FOR 499 S.A. Kotz).

sistent in response to voices from young speakers? Similarly, building on the widely held (but seldom confirmed) belief that females and males differ in their 'emotionality'(see [10]) , we aimed to specify if there are also gender voice specific differences in on-line emotional speech perception.

2. Methods

2.1. Participants

Thirty-two native speakers of German (sixteen female, mean age: 24.4 years; range 21-29 years) participated in the experiment. Participants were right-handed, had no reported hearing or neurological problems, and received financial compensation for their participation.

2.2. Stimulus Material

The material consisted of semantically and prosodically matching stimuli for each of four basic emotions (anger, fear, disgust, sadness) and a neutral baseline. For each emotion and sentence type, 20 sentences were presented, adding up to 80 sentences. These sentences were spoken by four different speakers (female/young, male/young, female middle-aged/ male/middle-aged), that is 320 emotional sentences were presented in total. In addition, 40 semantically and prosodically neutral sentences were presented. As each of the neutral sentences was also spoken by all four speakers, 160 neutral sentences were presented. In total, we presented 480 sentences in one experimental session. Emotional prosodic valence was obtained in an earlier rating study (for stimulus details, see Ref. [9]). The mean accuracy rates of the critical sentences presented are as follows: anger: 87%, fear: 59%, disgust: 69%, sadness: 63% neutral: 91% (chance level was 14%). All sentences were taped with a videocamcorder and later digitized at 16-bit/44.1 kHz sampling rate. The stimulus material was prosodically analyzed (i.e. pitch, intensity and duration of the sentences were extracted) using *Praat*. Results of acoustical analyses can be found in Table 1.

2.3. Procedure

Each participant was seated comfortably at a distance of 115 cm from a computer monitor in a sound-attenuating room equipped with a three-button response panel, with only the left and right button being relevant for the task. Half of the participants pressed the yes-button with their right hand and the no-button with their left hand. The sentences were presented via loudspeaker at a comfortable hearing level. Instructions with examples asked participants to listen to the presented sentence, read a following word (flashed on the screen for 300 ms) and to verify a probe as accurately and as quickly as possible (response answer time limit was set at 1500 ms). The intertrial interval was 1500 ms. Participants were asked to avoid eye movements during sentence presentation.

2.4. ERP Recording and Data Analysis

The electroencephalogram (EEG) was recorded with 58 Ag-AgCl electrodes mounted in an elastic cap according to the 10-20 system each referred to the nose (NZ). Bipolar horizontal and vertical EOGs were recorded for artifact rejection purposes. Electrode resistance was kept under 5kΩ. Data was re-referenced off-line to linked mastoids. The signals were recorded continuously with a band pass between DC and 70 Hz and digitized at a rate of 250 Hz. ERPs were filtered off-line with a 17 Hz low pass for graphical display, but all statistical analyses were computed on non-filtered data. Electroencephalogram recordings were scanned for artefacts. Separate ERPs for each condition at each electrode site were averaged for each participant with a 100-ms pre-stimulus baseline.

ERP components of interest were determined by visual inspection. For statistical analysis electrodes were grouped into six *Scalp Regions of Interest (SROI)*. Each following *SROI* defined a critical region of six scalp sites: left frontal (LF): F3, F5, F7, FC3, FC5, FT7; right frontal (RF): F4, F6, F8, FC4, FC6, FT8; left central (LC): C3, C5, T7, CP3, CP5, TP7; right central (RC): C4, C6, T8, CP4, CP6, TP8; left posterior: P3, P5, P7, PO3, PO7, O1; right posterior: P4, P6, P8, PO4, PO8, O2.

3. Results

Overall comprehension of the sentences was very good (overall accuracy score: 97%). Note, that behavioral responses are not reported because the implicit task was solely used to ensure that participants listened attentively to the sentences. For the ERP analysis, a repeated analysis of variance (ANOVA) was conducted in the time window between 150 - 300 ms. The time-window was based on previous evidence [8]. Analyses on ERP mean amplitudes for correctly answered trials were analyzed for the factors *speaker gender* (female or male), *speaker age* (young or middle-aged), *emotional expression* (anger, disgust, fear, sadness, and neutrality), and the topographical factors *hemisphere* (right/left hemisphere) and *region* (anterior/central/posterior region). Only significant interactions with critical factors (emotional expression, speaker gender, speaker age) are reported in step-down analyses. The null-hypothesis was rejected for *p*-values smaller than 0.05. The Geisser-Greenhouse correction was applied to all repeated measures with greater than one degree of freedom in the numerator. The p-values for break-down comparisons were corrected using a modified Bonferroni procedure [11]. In addition, effect size was estimated by omega-squared (ω^2).

3.1. ERP results

P200: In the time window of 150 ms to 300 ms, a significant effect of *emotional expression* was found ($F(1,31)=17.90$, $p<.0001$, $\omega^2 = 0.2552$) revealing waveform differences between emotional sentences. Break-down analyses confirmed that neutral sentences differed significantly from all emotional sentences. Results are listed in the following: 1) neutral vs. angry sentences ($F(1,31)=5.76$, $p<.05$); 2) neutral vs. disgust sentences ($F(1,31)=9.71$, $p<.01$); 3) neutral vs. fearful sentences ($F(1,31)=37.78$, $p<.0001$); 4) neutral vs. sad sentences ($F(1,31)=37.56$, $p<.0001$); with all comparisons showing more positive ERP waveforms for neutral sentences than for emotional sentences.

Also, *emotional expression* interacted with the factor *speaker gender* ($F(4,124)=5.60$, $p<.001$), suggesting different effects of emotional expression for female and male speakers. The step-down analysis by *speaker gender* revealed (marginal) significant emotional expression effects for both female ($F(4,124)=21.09$, $p<.0001$; $\omega^2 = 0.3010$) and male speakers ($F(4,124)=2.50$, $p=.06$, $\omega^2 = 0.0307$); however, effect sizes indicated a stronger effect for female speakers. The statistical values for the emotion effects of post-hoc comparisons are as follows. Female speakers: 1) neutral vs. fearful sentences ($F(1,31)=46.26$, $p<.0001$); 2) neutral vs. sad sentences ($F(1,31)=41.81$, $p<.0001$). Male speakers:. 1) neutral vs.

Speaker	Parameter	Emotion Anger	Disgust	Fear	Sadness	Neutrality
Young Female	Mean F0	282.5 (14.28)	221.9 (9.8)	244.0 (8.5)	259.7 (47.6)	222.2 (9.0)
	Duration	2.4 (0.2)	2.7 (0.3)	2.6 (0.1)	2.5 (0.2)	2.4 (0.2)
	Mean Intensity	68.0 (1.6)	63.0 (1.8)	62.6 (2.3)	64.8 (2.2)	65.1 (2.4)
Young Male	Mean F0	251.9 (27.75)	128.8 (31.8)	120.3 (9.6)	124.4 (14.5)	119.6 (8.3)
	Duration	3.0 (3.1)	2.9 (0.31)	4.2 (1.2)	2.9 (0.3)	2.8 (0.3)
	Mean Intensity	68.4 (1.6)	65.3 (2.8)	66.8 (3.0)	66.8 (1.2)	66.2 (2.9)
Middle-aged Female	Mean F0	279.2 (33.51)	248.6 (37.9)	239.4 (24.7)	190.2 (9.3)	196.7 (9.2)
	Duration	2.9 (0.2)	3.9 (0.4)	3.5 (0.4)	3.5(0.4)	3.4 (0.3)
	Mean Intensity	64.7 (1.4)	68.0 (2.1)	65.7 (2.2)	64.4 (2.2)	65.7 (2.0)
Middle-aged Male	Mean F0	181.1 (24.18)	136.8 (15.11)	199.4 (19.2)	119.4 (8.6)	108.2 (5.4)
	Duration	3.1 (0.4)	3.3 (0.4)	2.9 (0.3)	2.7 (0.3)	2.9 (0.3)
	Mean Intensity	63.5 (2.1)	62.5 (1.8)	64.5 (1.3)	64.8 (2.2)	62.3 (1.8)
All	Mean F0	248.7 (47.1)	184.0 (60.2)	200.8 (57.3)	173.5 (65.9)	161.7 (56.3)
	Duration	2.8 (0.3)	3.2 (0.5)	3.3 (0.7)	2.9 (0.4)	2.9 (0.4)
	Mean Intensity	66.1 (2.4)	64.7 (2.5)	64.9 (1.8)	65.0 (1.3)	64.8 (1.7)

Table 1: *Acoustical analyses for the four speakers presented in the experiment.*

disgust sentences (F(1,31)=7.04, $p<.05$); 2) neutral vs. fearful sentences (F(1,31)=3.92, $p=.06$); 3) neutral vs. sad sentences (F(1,31)=8.57, $p<.01$). In all contrasts neutral sentences elicited stronger P200 amplitudes than emotional sentences.

In addition, *emotional expression* interacted with the factor *speaker age* (F(4,124)=6.88, $p<.001$), revealing different effects of *emotional expression* between young and middle-aged speakers. The step-down analysis by *speaker age* revealed significant emotional expression effects for both young (F(4,124)=12.25, $p<.0001$, $\omega^2 = 0.1775$) and middle-aged speakers (F(4,124)=11.95, $p<.0001$, $\omega^2 = 0.1959$). The statistical values for the emotional expression effects of post-hoc comparisons are as follows. Young speakers: 1) neutral vs. fearful sentences (F(1,31)=17.78, $p<.001$) and 2) neutral vs. sad sentences (F(1,31)=6.38, $p<.05$). Middle-aged speakers:. 1) neutral vs. angry sentences (F(1,31)=20.41, $p<.0001$); 2) neutral vs. disgust sentences (F(1,31)=28.24, $p<.0001$); 3) neutral vs. fearful sentences (F(1,31)=17.59, $p<.001$); 4) neutral vs. sad sentences (F(1,31)=35.91, $p<.0001$). In all comparisons neutral sentences elicited stronger P200 amplitudes than emotional sentences.

Moreover, there was a three-way interaction of all three factors *emotional expression* x *speaker gender* x *speaker age* (F(4,124)=4.63, $p<.01$). This interaction was first resolved by *speaker age* and then by *speaker gender*. For each speaker, the *emotional expression* effect reached significance: young female speaker (F(4,124)=12.81, $p<.0001$, $\omega^2 = 0.2079$), young male speaker (F(4,124)=4.77, $p<.01$, $\omega^2 = 0.0741$), middle-aged female speaker (F(4,124)=11.65, $p<.0001$, $\omega^2 = 0.2028$), middle-aged male speaker (F(4,124)=4.75, $p<.01$, $\omega^2 = 0.0767$). P200 amplitudes in response to sentences spoken by the young female speaker differed for the following comparisons: 1) neutral vs. disgust (F(1,31)=4.20, $p<.05$), 2) neutral vs. fearful sentences (F(1,31)=16.40, $p<.001$), and 3) neutral vs. sad sentences (F(1,31)=8.44, $p<.01$). Sentences spoken by the young male speaker differed for the comparison neutral vs. fearful sentences (F(1,31)=7.32, $p<.05$). For the middle-aged female speaker, the post-hoc comparisons revealed the following effects: 1) neutral vs. angry sentences (F(1,31)=4.92, $p<.05$); 2) neutral vs. disgust sentences (F(1,31)=23.16, $p<.0001$); 3) neutral vs. fearful sentences (F(1,31)=31.33, $p<.001$); 4) neutral vs. sad sentences (F(1,31)=28.70, $p<.0001$). Last, post-hoc comparisons for the middle-aged male speaker revealed the following: 1) neutral vs. angry sentences (F(1,31)=12.60, $p<.01$) and 2) neutral vs. sad sentences (F(1,31)=9.28, $p<.01$)

The interactions *speaker age* x *region* (F(2,62)=3.77, $p=.05$) and *speaker gender* x *hemisphere* x *region* (F(2,62)=3.46, $p<.05$) reached significance, but step-down analyses did not reveal any significant effects.

Taken together, the results revealed a significant main effect of emotional expression. ERP amplitudes of the P200 component were more positive-going for neutral sentences than for all emotional sentences. In addition to interactions between emotional expression and each speaker identity feature investigated, there was a significant three-way interaction of emotional expression, speaker gender and speaker age, that suggests individual speaker differences for the emotional expression effect. Effects are visualized in Figures 1, 2, and 3.

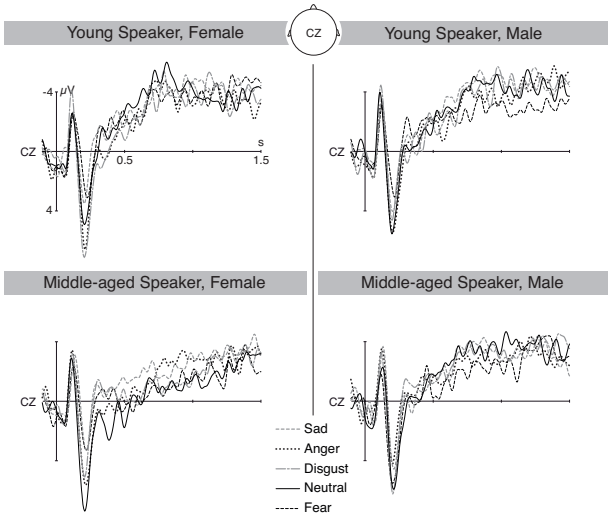

Figure 1: *P200 effect at one selected electrode site (CZ) for each individual speaker. Waveforms show the average for neutral (solid) and emotional (not solid) sentences from 100 ms before stimulus onset up to 1500 ms after stimulus onset.*

4. Discussion

The present investigation aimed to further specify the time-course of neural responses to early differentiation of vocal emotional and neutral expression, speaker identity, and their possible interplay. The results substantiate ERP evidence on the processing of vocal emotional expressions. The obtained effects are comparable to previous results [8] but extend these to four different speaker voices from two different age cohorts. The fact that basic vocal emotional expressions are differentiated from vocal neutral expressions in the P200 component for all four speakers points to the fact that early emotional differentiation is a highly robust phenomenon. In fact, the current results clearly show that irrespective of speaker voice gender or speaker voice age emotional vocalizations can be distinguished from neutral vocalizations very rapidly and thus complement and extend previous evidence.

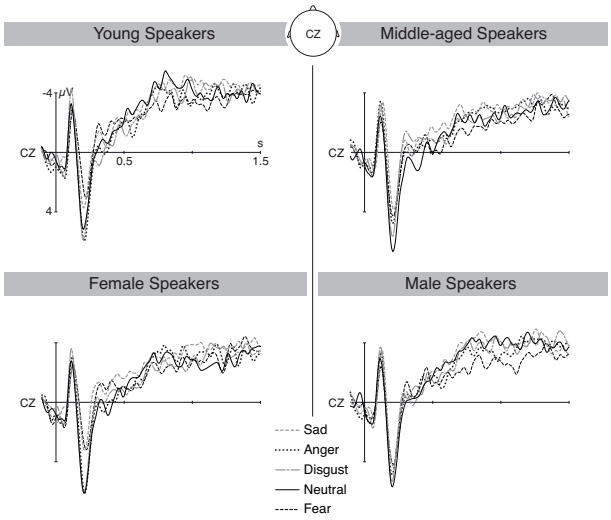

Figure 2: *P200 effect at one selected electrode site (CZ) for each speaker age and speaker gender group. Waveforms show the average for neutral (solid) and emotional (not solid) sentences from 100 ms before stimulus onset up to 1500 ms after stimulus onset.*

During experimental stimulation, participants were engaged in a probe verification task, thereby not explicitly focusing on the emotion or the speaker of the sentences. Thus, observed effects can not be related to attentional aspects, that is a first rapid emotional encoding occurs without focusing the attention on emotional aspects of the stimuli. It seems reasonable to suggest that this initial encoding is linked to emotional salience detection, while true emotional recognition of the stimulus may occur at a later point in time (c.f. [8]). In fact, results are very much in line with a recently proposed model on emotional prosody processing [12]. Within this model, a first structural encoding of an auditory stimulus is proceeded by the integration of acoustical stimulus features before more deliberate analyses of the stimuli occur. More specifically, the first two stages of emotional vocalization processing are supposed to take place within the first 200 ms after stimulus onset. Arguably, a first emotional encoding is primarily based on acoustical properties that may form an auditory object as early as 200 ms after stimulus onset. Given that emotional expressions seem to have their own individual acoustical configuration pattern (see e.g. [9]), one can assume that the observed early emotional differentiation, or salience detection, is not based on one single acoustic cue. Whether or not this early differentiation includes an emotional tagging process, that is an early assignment to one specific emotional category, needs to be explored in future research.

An additional aim of the current study was to investigate to which extent speaker identity influences vocal emotional processing. While previous investigations (e.g. [8, 13] report ERP differentiation for two speaker voices, the current experiment, including four different speaker voices, failed to find a significant speaker identity effect. Although no clear evidence for 'privileged' processing of female voices is observed, we report different effect sizes for female and male voices with regard to the emotion expression effect (with females showing a larger effect than male voices). This may again stir up the debate of whether female voices are more salient than male voices leading to clearer emotional signals in an utterance. An assumption, some may argue, which gets support from the observation that we fail to report such effect size differences with regard to speaker age (in which one female and one male voice of each gender group are collapsed). However, given the reported main effect of emotional expression as well as high emotional speech recognition rates for all four speakers in an earlier rating study [9], it seems unlikely that female voices induce 'privileged' processing in emotional speech processing. Rather, different acoustical profiles (e.g. high pitch vs. low pitch, fast vs. slow tempo, high vs. low intensity) irrespective of gender (or age) may initiate minor processing differences. This assumption gets additional support from the observation that each speaker voice elicited slightly different P200 patterns. For instance, for the young female speaker, the differentiation between neutral and sad, neutral and disgust, and neutral and fear vocalizations was most pronounced while for the middle-aged male speaker, the differentiation between neutral and angry and neutral and sad vocalizations was most pronounced. In contrast, no such peculiarity was observed for the female middle-aged speaker, that is all different emotional expressions tested differed from neutral vocalizations with respect to P200 amplitude size. Taken together, we conclude that individual capacity (irrespective of speaker gender or speaker age) to *encode* emotional vocalizations can lead to differently pronounced P200 amplitudes for the *decoders* of these utterances though this is not to imply that early emotional encoding is speaker identity dependent. In fact, the current data can be taken as support for the notion that vocal affect analysis and vocal identity analysis follow a similar time-course and that the two processes interact within the first 200 ms.

5. Conclusions

To summarize, the present findings are highly comparable to effects observed in previous studies [8], but results were extended by introducing more speaker voices. We aimed to answer if speaker identity and emotional analyses interact in time and if so when. The data suggest that a first interaction of these information types occurs within 200 ms after stimulus onset. Furthermore, in line with models on emotional prosody processing [12], we propose that acoustical patterns specific to emotional vocalizations drive early differentiation of emotional and neutral vocalizations. Results revealed that this first emotional encoding is a very rapid process not requiring attention on the emotional attributes of the stimulus, nor is it limited to specific speaker identities (e.g. young or female voices). As this emotional (or salience) detection occurs within the first 200 ms af-

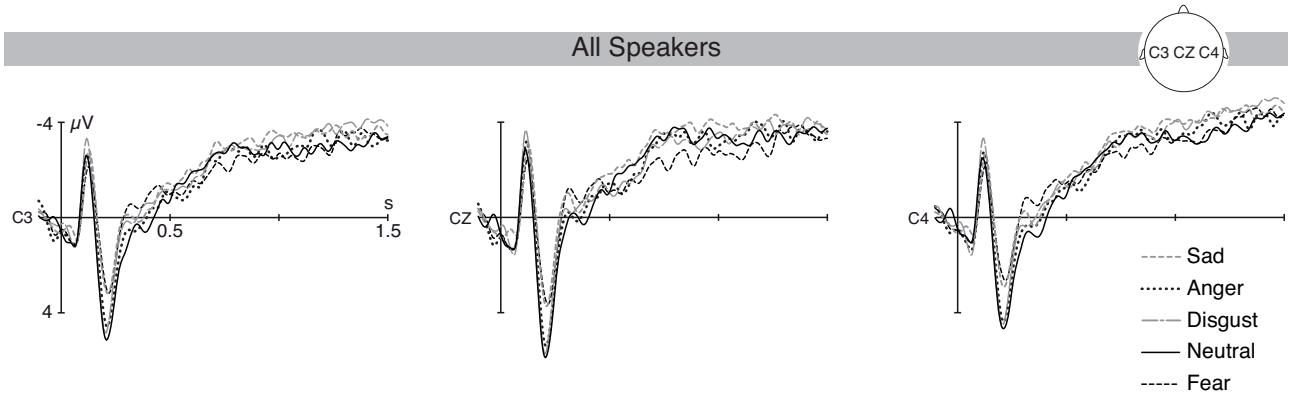

Figure 3: *P200 effect at selected electrode sites averaged across speaker. Waveforms show the average for neutral (solid) and emotional (not solid) sentences from 100 ms before stimulus onset up to 1500 ms after stimulus onset.*

ter stimulus onset, one should keep in mind that such an initial evaluation may need to be re-evaluated at a later point in time, that is, when more information is revealed. This, together with specifying whether the first differentiation is accompanied by a valence or emotional category tagging process, should be tested in future studies.

6. References

[1] Belin, P., Fectau, S., Bédard, C. 2004. Thinking the voice: neural correlates of voice perception. *TRENDS in Cognitive Sciences*, 8(3), 129-135.

[2] Hickok, G., Poeppel, D. 2000. Towards a functional neuroanatomy of speech perception. *TRENDS in Cognitive Sciences*, 4, 131-138.

[3] Zatorre, R.J., Binder, J.R. 2000. Functional and structural imaging of the human auditory system. *Brain Mapping: The Systems*, pp. 837-846. Academic Press.

[4] Scott, S.K., Blank. C.C., Rosen, S., Wise, R.J.S. 2000. Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123, 2400 - 2406.

[5] Mullenix, J. 1997. On the nature of perceptual adjustments to voice. In: Johnson, K., Mullenix, J., (eds.). *Talker variability in speech processing*. Sand Diego: Academic Press, pp. 67-83.

[6] Titova, N., Näätänen, R. 2001. Preattentive voice discrimination by the human brain as indexed by mismatch negativity *Neuroscience Letters*, 308, 63-65.

[7] Knösche, T.R., Lattner, S., Maess, B., Schauer, M., Friederici, A.D. 2002. Early parallel processing of auditory word and voice information. *NeuroImage*, 17, 1493-1503.

[8] Paulmann, S., Kotz, S.A. 2008. Early emotional prosody perception based on different speaker voices. *NeuroReport*, 19(2), 209-213.

[9] Paulmann, S., Pell, M.D., Kotz, S.A. in press. How aging affects the recognition of emotional speech. *Brain and Language*.

[10] Barrett, L. F., Robin, L., Pietromonaco, P. R., Eyssell, K. M. 1998. Are women the "more emotional sex?" Evidence from emotional experiences in social context. *Cognition and Emotion*, 12, 555-578.

[11] Keppel, G. 1991. *Design and analysis: a researcher's handbook*. Englewood Cliffs, NJ: Prentice Hall.

[12] Schirmer, A., Kotz, S.A. 2006. Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, 10, 24-30.

[13] Paulmann, S. Kotz, S.A. in press. An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain and Language*.