Multimodal processing of emotional meanings: A hypothesis on the adaptive value of prosody

PIERA FILIPPI[1, 2], SEBASTIAN OCKLENBURG[3], DAN BOWLING[4], LARISSA HEEGE[5], ALBERT NEWEN[2, 6], ONUR GÜNTÜRKÜN[2, 3], BART DE BOER[1]

[1]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium; [2]Center for Mind, Brain and Cognitive Evolution; Email: pie.filippi@gmail.com; [3]Department of Biopsychology, Ruhr University, Bochum, Germany, [4]Department of Cognitive Biology, University of Vienna, Austria; [5]Department of General and Biological Psychology, University of Wuppertal, Germany, [6]Department of Philosophy II, Ruhr University, Bochum, Germany.

Humans combine multiple sources of information to comprehend meanings. These sources can be characterized as linguistic (i.e., lexical units and/or sentences) or paralinguistic (e.g. body posture, facial expression, voice intonation, pragmatic context). Emotion communication is a special case in which linguistic and paralinguistic dimensions can *simultaneously* denote the same, or multiple incongruous referential meanings. Think, for instance, about when someone says "I'm sad!", but does so with happy intonation and a happy facial expression. Here, the communicative channels express very specific (although conflicting) emotional states as denotations. In such cases of intermodal incongruence, are we involuntarily biased to respond to information in one channel over the other? We hypothesize that humans are involuntary biased to respond to prosody over verbal content and facial expression, since the ability to communicate socially relevant information such as basic emotional states through prosodic modulation of the voice might have provided early hominins with an adaptive advantage that preceded the emergence of segmental speech (Darwin 1871; Mithen, 2005). To address this hypothesis, we examined the interaction between multiple communicative channels in recruiting attentional resources, within a Stroop interference task (i.e. a task in which different channels give conflicting information; Stroop, 1935). In experiment 1, we used synonyms of "happy" and "sad" spoken with happy and sad prosody. Participants were asked to identify the emotion expressed by the verbal content while ignoring prosody (Word task) or vice versa (Prosody task). Participants responded faster and more accurately in the Prosody task. Within the Word task, incongruent stimuli were responded to more slowly and less accurately than congruent stimuli. In experiment 2, we adopted synonyms of "happy" and "sad" spoken in happy and sad prosody, while a happy or sad face was displayed. Participants were asked to identify the emotion expressed by the verbal content

while ignoring prosody and face (Word task), to identify the emotion expressed by prosody while ignoring verbal content and face (Prosody task), or to identify the emotion expressed by the face while ignoring prosody and verbal content (Face task). Participants responded faster in the Face task and less accurately when the two non-focused channels were expressing an emotion that was incongruent with the focused one, as compared with the condition where all the channels were congruent. In addition, in the Word task, accuracy was lower when prosody was incongruent to verbal content and face, as compared with the condition where all the channels were congruent. Our data suggest that prosody interferes with emotion word processing, eliciting automatic responses even when conflicting with both verbal content and facial expressions at the same time. In contrast, although processed significantly faster than prosody and verbal content, faces alone are not sufficient to interfere in emotion processing within a three-dimensional Stroop task. Our findings align with the hypothesis that the ability to communicate emotions through prosodic modulation of the voice – which seems to be dominant over verbal content - is evolutionary older than the emergence of segmental articulation (Mithen, 2005; Fitch, 2010). This hypothesis fits with quantitative data suggesting that prosody has a vital role in the perception of well-formed words (Johnson & Jusczyk, 2001), in the ability to map sounds to referential meanings (Filippi et al., 2014), and in syntactic disambiguation (Soderstrom et al., 2003). This research could complement studies on iconic communication within visual and auditory domains, providing new insights for models of language evolution. Further work aimed at how emotional cues from different modalities are simultaneously integrated will improve our understanding of how humans interpret multimodal emotional meanings in real life interactions.

References

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London: John Murray.

Filippi, P., Gingras, B., & Fitch, W. T. (2014). Pitch enhancement facilitates word learning across visual contexts. *Frontiers in Psychology*, *5*, 1-8.

Fitch, W. T. (2010). *The evolution of language*. Cambridge University Press.

Johnson, E. K., & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language*, *44*, 548–567.

Mithen, S. J. (2005). *The singing Neanderthals: The origins of music, language, mind, and body*. London: Weidenfeld Nicolson.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science*, *33*, 127-146.

Rigoulot, S., & Pell, M. D. (2012). Seeing Emotion with Your Ears: Emotional Prosody Implicitly Guides Visual Attention to Faces. *PloS One*, *7*(1), e30740.

Soderstrom, M., Seidl, A., Nelson, D., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, 49, 249-267.

Stroop, J. R. (1935). Studies of Interference in Serial Verbal Reactions, *Journal of Experimental Psychology*, *18*, 643-662.