BRIEF REPORT

Compatibility Between Tones, Head Movements, and Facial Expressions

Gernot Horstmann
Bielefeld University

Ulrich Ansorge
University of Vienna

The study tests the hypothesis of an embodied associative triangle among relative tone pitch (i.e., high or low tones), vertical movement, and facial emotion. In particular, it is tested whether relative pitch automatically activates facial expressions of happiness and anger as well as vertical head movements. Results show robust congruency effects: happiness expressions and upward head tilts are imitated faster when paired with high rather than low tones, while anger expressions and downward head tilts are imitated faster when paired with low rather than high tones. The results add to the growing evidence favoring an embodiment account that emphasizes multimodal representations as the basis of cognition, emotion, and action.

Keywords: emotion, expression, embodiment

Understanding social-communicative behavior is a key concern in psychology. In communication, nonverbal signals such as facial expressions and gestures are of particular interest. In common belief, nonverbal signals can be more veridical of what social partners believe, feel, or intend than verbal communication (e.g., Ekman, 1972; but see Fridlund, 1994). However, apart from reflecting inner states such as beliefs or feelings, nonverbal signals also depend on situational cues, including the presence of a relevant social audience (Chovil, 1991; Kraut & Johnston, 1979), the observation of another person experiencing pain (Bavelas, Black, Lernemy, & Mullett, 1986), the exposure to an expressive pain (Dimberg, 1982), or the presence of low or high tones (Horstmann, 2010). In many cases, nonverbal signals and situational cues appear to form sensory-motor couplings which are automatic in the classical sense (cf., Posner & Snyder, 1975), that is, in being triggered without intention, proceeding without awareness, and without tight volitional control.

In most accounts, situational cuing either directly fulfills communicative functions or appears as a side effect of otherwise adaptive principles. For example, wincing while observing another person’s pain signals a basic form of empathy, communicating “I feel as you feel” (Bavelas et al., 1986); and the automatic mimicry of facial expressions has been interpreted as a side effect of the sensory-motor loops involved in the perception of facial expressions (Niedenthal, Brauer, Halberstadt, & Innes-Ker, 2001).

Sensory-motor couplings in nonverbal behavior have become a key finding for an embodied cognition and emotion theoretical view. This embodiment approach (e.g., Barsalou, 2008; Balcetis & Cole, 2009; Glenberg & Kaschak, 2003; Niedenthal, Barsalou, Winkielman, Krauth-Gruber, & Ric, 2005) opposes the traditional cognitive science view which emphasizes symbolic information processing and propositional representations within a highly modular architecture of the mind (e.g., Carruthers, 2005; Fodor, 1983; Pylyshyn, 1984). In contrast to this, the embodiment approach emphasizes analogical modality-based information processing and a highly interconnected architecture of the mind. The modal representations of a smile, for example, comprise sensations, perceptions, bodily feelings, and actions that usually co-occur during episodes of smiling. The modal components, in turn, are connected via supramodal convergence zones (Damasio, 1989) which allow their reciprocal activation. Because of intermodal connections, cues from one modality can automatically activate components of a representation in other modalities.

An Intermodal Associative Triangle of Relative Tone Pitch, Facial Expression, and Vertical Direction

The present work deals with sensory-motor couplings that form an intermodal associative triangle between relative tone pitch, facial expression, and vertical direction. As to the pitch-expression association, converging evidence from different disciplines within and outside psychology shows that the distinction between relatively high versus low tones is connected to signals of friendliness and happiness, versus to signals of threat and anger. For example, a smile can be heard in concomitant vocalizations because of a relatively high fundamental frequency (F0; Drahota, Costall, & Reddy, 2007). F0 is the slowest sine-wave oscillation of a sound, corresponding roughly to the phenomenally perceived pitch. Research on vocal expression of emotion has found high F0 tied to happiness (e.g., Justlin & Laukka, 2003; Sauter, Eisner, Calder, & Scott, 2010), and developmental research found that infants show positive facial expressions when praised in a high voice and negative expressions when scolded in a low voice (Fernald, 1993). Additionally, Morton (1977, 1994) analyzed vocal behavior from 54 species and found that low pitched sounds signal aggression, whereas high pitched sounds signal helplessness, friendliness, and submission. Second, tone pitch has strong spatial associ-
ations with vertical space. For example, Ben-Artzi and Marks (1995) revealed faster categorization of vertical position accompanied by irrelevant tones in compatible trials (high positions and high tones, low positions and low tones) than in incompatible trials (see also, Walker & Smith, 1984). Finally, as to the association between vertical direction and emotional expressions, we first note that social gestures are inherently spatial, as any physical action is. Accordingly, theorists of signaling have interpreted nonverbal signals as intention movements (Heinroth, 1911; Fridlund, 1991, 1994) revealing the initial phase of a forthcoming action (e.g., teeth baring being an intentional movement signaling an attack which might include biting). With particular relevance to vertical direction and emotion, facial expressions of happiness and anger entail upward movements of the lip corners in smiling (and sometimes also of the eyebrows as in the eye-brow flash; cf. Eibl-Eibesfeldt, 1972), while the expression of anger typically entails a downward movement of the lip corners and the inner brows. Additionally, according to a study by Rosenberg and Langer (1965), extreme anger may be expressed by a combined forward/downward movement of the whole body.

Aim of the Present Study

The present study seeks further insight into the embodied associative triangle between facial expression, relative tone pitch, and vertical direction. The sensory-motor associations between tone pitch and facial expressions are reflected in the tone-affect compatibility effect (Horstmann, in press): When participants imitated facial expressions of emotion that were presented simultaneously with short (100 ms) tones, happy facial expressions were imitated faster when the tone was high relative to low, whereas angry facial expressions were imitated faster when the tone was low relative to high. The present study seeks to (a) replicate this effect with a modified procedure that allows more reliable conclusions about the automaticity of the process, and to (b) probe the second part of the multimodal triangle: the association between pitch and vertical body movements.

As to the first aim, Horstmann (in press) used a dual-task paradigm where the tones had to be categorized after the facial response. Thus, tone discrimination was part of the task, undermining the inference that the congruency effect was independent from intention and in this sense automatic. In the present Experiment 1, the tone was completely task-irrelevant and participants were instructed to ignore it. In addition, Experiment 2 used a task where attention is diverted from tone pitch, to further test the limits of automaticity of the effect.

As to the second aim, the test of an embodied association between pitch and vertical movement, Experiment 1 tests two conditions (see Figure 1). In the face condition, the task closely resembled the one in Horstmann (in press, Experiment 1): Participants had to imitate a happy or an angry face, which was accompanied by a high or low tone. The head condition was structurally the same as the face condition, with the only critical change being that the to-be-imitated movement was an upward versus a downward head tilt. The head condition specifically tests the hypothesis that pitch primes vertical head movements—we expected faster (or more accurate) downward head tilts to low than to high tones and faster (or more accurate) upward head tilts to high than to low tones.

![Figure 1. Results for Experiment 1. Error bars show the standard error of the mean.](image)

**Experiment 1**

**Method**

**Participants.** Participants were 24 students, 12 women (age: $M = 26.8$, $SD = 4.8$). Some were psychology students who received course credits; others, recruited in the Bielefeld University public hall, received candy as a symbolic gratification.

**Stimuli.** Tones were monophonic, 100-ms square waves of 300 or 600 Hz (cf. Horstmann, in press), corresponding to a relatively low or high tone, respectively. They were presented via the internal speaker of the computer. Sound level was about 50 dB(A), as measured at the speaker, and far less (below the measuring range of a Noris NM-3 sound level meter) at the participant’s ear in about 150 cm distance. The color photographs were custom-made for this experiment (see Figure 2). They depict a woman showing closed-mouth expressions of happiness or anger, as well as a left or a right head turn, and an upward or downward head tilt. The expressions corresponded to Ekman and Friesen’s (1978) criteria for closed mouth expressions of happiness and anger, respectively. For the head movements, expression was neutral. The faces covered the entire 15-in. screen and were thus almost the size of real faces. A single woman’s face was used to reduce error variance in the reaction times (RTs).

**Apparatus.** A Pentium II computer controlled stimulus presentation. Facial responses were recorded with a SONY camcorder, positioned behind the monitor. A mirror on a tripod behind the participant reflected the screen, such that the screen and the participant’s face were simultaneously videotaped.

**Procedure.** Participants were instructed to rapidly and accurately imitate the facial expression on the screen. The tone was not task relevant; participants were informed that tones were paired equally often with all faces, and thus provided no information about the upcoming face. Half of the participants were assigned to the head condition, and half to the face condition. Each condition comprised 12 practice trials, followed by four blocks, each with 32 experimental trials, separated by breaks. Each trial began with the instruction “Neutral face!” for 2 s to ensure that the participant’s face was neutral at the beginning of the next trial, followed by a fixation cross for 1 s. Next the tone and the picture started simultaneously. The tone and picture durations were 100 ms and 2 s, respectively. The picture was followed by a blank screen for up to 8 s. Participants terminated the blank screen and started the next trial with a key press.
All participants also completed a control condition in which no compatibility effect was expected. In the control condition, pictures showed the woman with her head turned to the left or to the right. Pitch again varied randomly. The orthogonal combination of head position and pitch resulted in four conditions that were replicated 16 times. Half of the participants completed this condition before and half after the experimental conditions. The control condition served as a test of a baseline difference between the groups.1

Data analysis. The main dependent variable was the latency of the facial or head response (Horstmann, in press). The video recordings were analyzed offline. Each trial was played back and forth until the onset video frames of the stimulus, and the onset of the correct response was identified. Only trials where the first response was correct were scored as correct. If no response was given, the trial counted as an error. One coder coded all trials of all participants. A second coder coded two randomly chosen participants to determine interrater reliabilities (.81 and .91, computed over 204 observations each). Videos were mute during coding such that coders were blind to the tone manipulation. The data from one participant had to be dropped because he left before the end of the experiment to keep an appointment.

Results

Facial RTs below 180 ms or exceeding 1,480 ms were discarded (<1%) to exclude prepotent and disorderly long responses from the analysis (note: cut-offs are multiples of 40 ms, which is the duration of one video frame). In the control task, mean correct RTs were analyzed with a 2 (group: head vs. face) × 2 (tone: high vs. low) × 2 (head movement: left vs. right) ANOVA, which revealed a main effect for head movement only, \( F(1, 21) = 9.59, p < .01 \), reflecting that the turn to the left was 6 ms slower than the turn to the right (292 vs. 298 ms). All other effects were clearly nonsignificant, \( Fs < 1.00 \). A corresponding error analysis showed no main effects or interaction, \( Fs < 1.00 \). Mean error rate was 3%.

Thus, the analysis of the control task revealed no differences for the between-participants variable condition and also no differences for the within-participants variable pitch. Errors are displayed in Table 1.

For the face condition, mean RT was analyzed with a 2 (tone: high vs. low) × 2 (expression: happy vs. angry) ANOVA. It revealed a Tone × Expression interaction only, \( F(1, 11) = 18.62, p = .001, \eta^2 = .63 \) (other \( Fs < 1.59 \)). Follow-up \( t \) tests revealed significant differences for both expressions, both \( ts > 2.7, ps < .01 \) (one-tailed). This was not attributable to a speed–accuracy trade-off: a corresponding ANOVA of the error scores revealed the same pattern of results \( F(1, 11) = 9.48, p = .010, \eta^2 = .46 \) (other \( Fs < 1 \)). Errors were higher in the incompatible conditions (1.6%) than in the compatible conditions (0.23%).

For the head condition, a corresponding 2 (tone: high vs. low) × 2 (movement: up vs. down) ANOVA revealed a Tone × Movement interaction only, \( F(1, 10) = 6.86, p = .025, \eta^2 = .41 \) (other \( Fs < 1 \)). Follow-up \( t \) tests revealed significant differences for both movement directions, \( t > 1.9, p < .04 \) (one-tailed). Again, the RT-pattern was not attributable to a speed–accuracy trade-off: a corresponding ANOVA of the error scores revealed a significant interaction only, \( F(1, 10) = 9.69, p = .011, \eta^2 = .49 \) (other \( Fs < 1 \)). Errors were higher in the incompatible conditions (2.7%) than in the compatible conditions (1.4%).

Discussion

Experiment 1 found a cross-modal compatibility effect in two conditions. With a high tone, happiness was imitated faster than
anger, whereas with a low tone, anger was imitated faster than happiness. Furthermore, with a high tone, an upward head movement was faster than the downward head movement, whereas with a low tone, a downward movement was faster than the upward movement.

The facial and head responses reported here are quick. The latencies are similar to visuo-manual responses when stimulus-response compatibility is high (e.g., Dutta & Proctor, 1992). This is in line with Dimberg and Thunberg (1998), who also found facial imitation to be fast (300–400 ms, as measured with EMG). Some studies (e.g., Moody, McIntosh, Mann, & Weisser, 2007) found somewhat higher latencies; there are, however, multiple differences in the tasks, each of which might have altered response latency.

**Experiment 2**

Experiment 1 showed that irrelevant tones in themselves, in the absence of a task to process them, prime particular gestures and facial expressions. However, was intentional processing of tonal height discouraged by making tones irrelevant in Experiment 1? While using the same primary task as in the face condition of Experiments 1, Experiment 2 used a second task to direct participants’ attention away from tone pitch toward a nonrealted feature of the tone: its left or right position. The requirement to focus on tone position should attenuate any tendency to willingly encode pitch. An automatic effect, however, should survive this manipulation.

**Method**

**Participants.** Fourteen students (13 women), with a mean age of 22.8 years took part for £2 payment. Two additional cases were tested but dropped because their facial responses lacked sufficient distinctness.

**Stimuli and procedure.** These differed in two aspects from the face condition in Experiment 1. Importantly, participants responded to tone location (the left or right through external speakers) by pressing the left key for the left tone and the right key for the right tone. The response to the tone was unspeeded and participants were clearly instructed to indicate tone location always as the second response, after they had imitated the face. Tone pitch and tone location were not correlated. As a minor change, the pair of pictures, showing open mouth displays of happiness and anger, was the same as in Horstmann (in press, Experiments 2 and 3).

**Results**

Mean correct facial RTs (see Figure 3) were analyzed with a 2 (tone: high vs. low) × 2 (expression: friendly vs. angry) × 2 (tone location: left vs. right) ANOVA, which revealed a significant Tone × Expression interaction, $F(1, 13) = 23.7, p < .001, \eta^2 = .64$, and a marginally significant tone effect, $F(1, 13) = 4.1, p = .06, \eta^2 = .25$. Smiles were imitated 27 ms faster with high than with low tones, $t(13) = 2.93, p < .05$, and frowns were imitated 44 ms faster with low than with high tones, $t(13) = 5.86, p < .001$. The other main effects and the interactions were not significant, all $Fs < 1.9$. Error rates were far too low (i.e., < 1.0%) to warrant further analysis.

**Discussion**

We found a clear congruency effect in a task in which willed processing of pitch should be strongly attenuated by the secondary task to report the position of the tone. Thus the present experiment supported the claim that the compatibility effect between pitch and facial expressions was attributable to automatic and not willed processing of pitch.

**General Discussion**

The present study revealed robust sensory-motor coupling of sounds and social gestures. Facial expressions of happiness and upward head tilts were faster after high than low tones, whereas facial expressions of anger and downward head tilts were faster after low than after high tones. Quicker responses in compatible than in incompatible conditions were observed in 20 of 23 participants in Experiment 1, and in 13 of 14 in Experiment 2, implying quite a robust effect.2

The present experiments also show that the sensory-motor coupling is highly automatic. In Experiment 1, the tones were completely task-irrelevant and participants were instructed to ignore them. In Experiment 2, tones were task-relevant, but tone pitch was to be ignored because it varied orthogonally to the task-relevant horizontal position of the tone.

For the first time, we found a sensory-motor coupling of relative pitch and vertical head movement. This result adds to the existing evidence for an embodied associative triangle between pitch, facial expression, and vertical movement in the form of direct evidence for associations between tone pitch and facial expressions and associations between tone pitch and vertical movement.

**Embodiment Versus Metaphor**

In the Introduction we discussed an embodiment approach. Embodiment motivated the present research particularly because of the assumption of multimodal representations which readily explains the automatic interaction between simple stimuli—tones—and motor responses—nonverbal behavior. However, the association between verticality and valence has also been exam-

---

2 Some readers might regard the sample sizes as rather small, raising doubts on the validity of the results. Note, however, that replication of effects in small sample sizes might be a better indicator of validity than large sample sizes (e.g., Cohen, 1994).
ined in research on conceptual metaphors (Lakoff & Johnson, 1980), which assumes that humans conceptualize abstract meaning metaphorically in terms of perceptual concepts, such as magnitude or space. In support of this account, Meier and colleagues (Meier & Robinson, 2004; Meier, Robinson, & Clore, 2004) found valence to be associated with lightness (light vs. dark), vertical position (high vs. low), and tone pitch (high vs. low). Thus one might argue that the present results can also be explained by means of conceptual metaphor.

However, while conceptual metaphor theory easily explains the association between semantic concepts of height and valence, it is not clear whether metaphorical meaning can explain the direct effects of physical stimuli (tones, positions) on actual movements (faces, head tilts). This is because the metaphor grounds conceptual meaning in physical dimensions; however, there is little requirement to metaphorically ground physical stimuli and physical action.

Relatedly, the metaphor account is particularly superfluous in understanding the meaning of emotions. Concepts, such as “law” or “news,” are abstract. Their meaning profits from and might even require grounding in fundamental perceptual attributes. Emotions such as happiness, pride, anger, or fear, however, are concrete. They can be used to ground abstract conceptual knowledge, but they need not be grounded themselves. Emotions are among the building blocks of cognition—like the perception of space, emotions are experiential priors. One might argue that emotions are fleeting, and perceptual grounding would be needed for this reason. However, this is true for virtually all perceptual phenomena, both in the realm of exteroception, such as movement, and interoception, such as thirst or hunger. It turns out that if we question the grounding of the perceptual priors, we are left devoid of any grounding principles. To summarize our position; while we agree that abstract concepts have no single perceptual grounding and thus need to be grounded metaphorically, emotions are grounds par excellence, being connected with distinctive feeling states, motivations, expression patterns, and physiology.

Embodied Multimodal Emotion Representations or Just Tone-Verticality Associations?

The present results are consistent with the use of multimodal embodied representations for nonverbal signaling. On this account, expressions of happiness and anger are differentiated in vocalization and facial expressions (Horstmann, in press). By the same token, signaling happiness entails upward movements (lip corners, eyebrows), whereas signaling anger entails downward movements (lip corners, jaw, brow). Consequently, high tones go together with upward movements and low tones with downward movements.

Alternatively, however, the present results are also compatible with another interpretation: tones prime the upward and downward movements entailed in emotional faces. This verticality hypothesis has the advantage of parsimony: It explains the present results with only one principle. The disadvantage of this hypothesis is that it does not (or only poorly) explain the established connection between pitch and emotion (see Introduction).

It might well be, however, that the typical approach in experimental psychology of dissecting causes is an unsuited oversimplification when applied to the unconfounding of tone-expression and tone-verticility associations. Embodied associations between gestures and vocalization must have evolved together during phylogeny. It is very likely that embodied associations between modalities developed as a means of redundant coding of the most important social communicative signals. Embodied associations persisted in humans as a means of nonverbal signaling, at least to secure reliable communication during preverbal phases of ontogenetic development. Under this perspective, telling apart the supremacy of tone-expression and tone-verticility associations takes the form of a chicken-and-egg problem.

Emotion Categories Versus Valence

In this paper, we endorse an emotion rather than a valence view. That is, we assume our effects to be specific for particular expression categories—anger and happiness—rather than for a dimension of valence. One reason is that this research originated from a social signaling perspective (cf. Fridlund, 1994), where nonverbal signals are assumed to have a more specific meaning than “I am good” or its negation. For example, the most prominent negative emotions, anger and sadness, may share the feature of being signals are assumed to have a more specific meaning than “I am good” or its negation. For example, the most prominent negative emotions, anger and sadness, may share the feature of being negative but are on opposite poles with respect to activity (with anger being high and sadness being low in activity; Russell, 1980) and with respect to the direction of behavior (approach vs. avoidance; Harmon-Jones, 2003).

The Emotional Meaning of Head Tilt

Head tilt is a gesture with social meaning and a frequently displayed nonverbal behavior. However, the emotional meanings of head tilt are manifold. For one, tilting the head up can be seen as a display of superiority, mastery, or dominance. For example, Zahavi and Zahavi (1997) interpret raising the chin as a costly superiority signal: the chin is vulnerable in a fight, and presenting it conveys self-assurance and scorn the offender. For another, tilting the head down can be seen as a display of despair (e.g., Stepper & Strack, 1993) but also as an intention movement (Heinroth, 1911), being part of an attack, simultaneously indicating the

Figure 3. Results for Experiment 2. Error bars show the standard error of the mean.
direction of movement of the attack (forward) and already assuming a defensive position (shielding the throat from blows). The association of a downward movement of the head with both defeat and attack supports Fridlund's (1994) conjecture that the social meaning of expressions is not fixed but depends on the context.

**Conclusion**

We present new evidence for an embodied associative triangle among relative pitch, vertical direction, and facial emotion. This associative triangle is inferred on the basis of automatic cuing of facial signals by relative pitch. Theoretically, the present results add to the growing evidence supporting an embodiment account that emphasizes multimodal representations as the basis of cognition, emotion, and action.

**References**


