Presubmission enquiry: Trends in Cognitive Sciences

Face and voice perception: commonalities and differences

To: Dr Lindsey Drayton, Editor, Trends in Cognitive Sciences.

23.07.19.

Dear Lindsey,

We are attaching a presubmission enquiry concerning a paper we would like to write for *Trends in Cognitive Sciences*.

Our topic is parallels and differences between the perception of faces and voices. A substantial body of modern research has focussed on communication from the face and voice, demonstrating notable parallels and thus leading to the important theoretical suggestion that the voice can be considered as a kind of 'auditory face'; for example, in an important TiCS review (Yovel & Belin, 2013, *TiCS, 17*, 263-271)

Nonetheless, this view has not gone unchallenged. For example, a recent meta-analysis (Schirmer, 2018, *SCAN, 13,* 1-13) has questioned this interpretation, calling for a more modality-specific perspective.

Our aim here is to therefore evaluate the ways in which the functional organisation of face and voice perception offers parallels, but placing these in the context of ways in which faces and voices also differ. Moreover, and of particular importance, we take the discussion to the next level by asking *why* these commonalities and differences exist? We offer a novel account grounded in the interaction between intrinsic characteristics of faces and voices and the demands of everyday life.

We hope that this agenda will be of interest to *TiCS* and its readers. We are confident that we are able to offer a novel theoretical perspective based on compelling studies to back up our case.

We look forward to hearing from you.

Yours sincerely,

Andy Young, Sascha Frühholz and Stefan Schweinberger.

Universities of York, Zurich and Jena.

Presubmission enquiry - ***Trends in Cognitive Sciences***

Authors: Andrew W. Young (University of York, UK), Sascha Frühholz (University of Zurich, Switzerland) and Stefan R. Schweinberger (University of Jena, Germany).

Face and voice perception: commonalities and differences

Human nonverbal communication involves complex patterns of signals communicated primarily through the face, voice, and body. Faces and voices in particular can often convey common forms of information concerning a person's gender, age, identity and emotional state, and they create impressions of warmth, competence and other social traits.

Most modern research has focussed on communication from the face and voice. Comparisons between face and voice perception have led to notable parallels and the important theoretical suggestion that the voice can be considered as a kind of 'auditory face' [1-3]. Nonetheless, this view has not gone unchallenged; a recent meta-analysis has questioned this interpretation, calling for a more modality-specific perspective [4].

Our aim here is therefore to evaluate the ways in which the functional organisation of face and voice perception offers such parallels and the ways in which they differ. Moreover, we take the discussion to the next level by asking *why* these commonalities and differences exist? We offer a novel account grounded in the interaction between intrinsic characteristics of faces and voices and the demands of everyday life.

Even beyond obvious differences in signal availability (for instance, the important role of facial eye gaze in signalling someone´s focus of attention has no direct counterpart in the voice), there are also important general differences between facial and vocal communication. In particular, the voice allows communication when the face is invisible, the voice can be silent even when the face is visible, nonverbal vocal communication often arises as a concomitant of speech, and in most everyday contexts a person can hear their own voice but can't see their own face.

We will develop a new theoretical perspective by reviewing understanding of the functional and neural organisation of nonverbal auditory communication in light of this overarching background, the different demands it creates, and the ways in which these act as determinants of a system that has to balance the needs of the sender and recipient.

Recognition of identity

We begin by considering the recognition of identity, which is often thought to form a paradigmatic example of the need to determine a relatively stable personal characteristic [5-8]. Whilst person identity can be determined from facial, vocal or body cues [1-3, 9-10], the existence of parallel forms of agnosic deficit (prosopagnosia and phonagnosia) strongly suggests a degree of modality-specificity in face and voice recognition mechanisms [10], and functional brain imaging studies implicate substantially different underlying brain regions. This modality-dependent separation may be driven by natural environments, in which a person's face is often seen before their voice is heard, or sometimes the voice is heard before the face is seen. Either circumstance puts a premium on a modality-specific mechanism that does not demand multimodal input [11].

Commonalities in the neural coding of faces and voices include phenomena that were first demonstrated for faces, but have later been observed for voices too, such as contrastive adaptation aftereffects [12-13]. Nonetheless, the neural coding principles underlying analysis and representation of faces and voices likely also differ. Face perception has to involve a strong degree of integration of spatially distributed information (eye, nose, mouth), whereas voice perception demands the integration of temporal information (acoustic information over time). That said, the frequent co-occurence of facial and vocal communicative signals does lead to a degree of cross-modal integration that is evident in some circumstances [14].

Importantly, there are other parallels between the functional demands of face and voice recognition - primarily involving recognition of familiar identities across substantial natural variation [6-7, 15-16] - that also drive corresponding forms of organisation in order to achieve recognition of familiar individuals across widely differing examples of the same face or the same voice. For this reason, recognition of familiar faces or voices far outstrips recognition of unfamiliar faces and voices [6-7, 16], and impairments in recognition of familiar faces or voices can occur even in the context of preserved recognition or matching of their unfamiliar counterparts.

Recognition of emotion

Having considered the factors that shape functional mechanisms underlying recognition of identity, we turn to recognition of emotion. Critically, emotions can change from moment to moment, and in consequence a strikingly different type of functional organisation arises for emotion recognition, where it has become evident that comprehension of facial and vocal cues is closely integrated [11, 17]. Patients with neuropsychological deficits following brain injury that affect emotion recognition invariably have problems that affect both facial and vocal expressions [11], and facial and vocal emotion recognition impairments also co-occur in other disorders including Parkinson´s disease and autism. Functional brain imaging points to a critical role of posterior STS for integrating audio-visual signals of emotion [11, 17-18]. We suggest that this form of organisation is driven by the need to resolve transient signals about mental and emotional states that require rapid readjustment of the perceiver's interpretation and intentions. Moreover, facial and vocal signals of emotion are themselves inherently somewhat ambiguous, but these ambiguities arise in different ways that often make the signals complementary. A multimodal mechanism that can integrate facial and vocal cues thus represents an optimal solution to these environmental and behavioural demands.

This type of multimodal organisation is especially clearly seen in the case of speech perception, where the temporal demands are even higher - requiring disambiguation of cues that may only last for milliseconds - and the nature of the complementarity between auditory cues and cues that can be read from movements of the lips and tongue is well established.

Nonetheless, besides the common role and functional importance of the STS in multimodal integration of facial and vocal cues to emotion, there are also some important differences in the functional neural network for emotional processing across modalities. While facial cues of emotions largely rely on a network including the visual cortex, posterior STS and the amygdala, vocal cues to emotion are processed by the mid and posterior STS and regions in the inferior frontal cortex [19-20].

Conclusions

We conclude that there are indeed parallels between face and voice perception that make the 'voice as an auditory face' metaphor a useful place to begin, but that it is time to appreciate strong differences between face and voice perception on the cognitive and neural level that are best understood as consequences of behavioural and environmental demands. Considering the complementary influences of intrinsic differences between faces and voices and the impact of the different demands of everyday life offers a novel perspective from which to understand the functional and neural organisation of how faces and voices are used in interpersonal perception.

References

1. Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences, 8*, 129-135.
2. Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*, 535-543.
3. Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences, 17*(6), 263-271.
4. Schirmer, A. (2018). Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing. *Social Cognitive and Affective Neuroscience, 13(1)*, 1-13.
5. Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*, 223-233.
6. Young, A. W., & Burton, A. M. (2017). Recognizing faces. *Current Directions in Psychological Science, 26*, 212-217.
7. Young, A.W., & Burton, A.M. (2018). Are we face experts? *Trends in Cognitive Sciences, 22,* 100-110.
8. Bernstein, M., & Yovel, G. (2015). Two neural pathways of face processing: a critical evaluation of current models. *Neuroscience and Biobehavioral Reviews, 55*, 536-546.

9. Yovel, G., & O'Toole, A. J. (2016). Recognizing people in motion. *Trends in Cognitive Sciences, 20(5)*, 383-395.
10. Barton, J. J. S., & Corrow, S. L. (2016). Recognizing and identifying people: a neuropsychological review. *Cortex, 75*,132-150.
11. Young, A.W. (2018). Faces, people and the brain: the 45th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology, 71*, 569-594.
12. Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*, 557-561.
13. Schweinberger, S. R. *et al.* (2008). Auditory adaptation in voice perception. *Current Biology, 18(9)*, 684-688.
14. Schweinberger, S. R., & Robertson, D. M. C. (2017). Audiovisual integration in familiar person recognition. *Visual Cognition, 25*, 589-610.
15. Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40,* 202–223.
16. Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin and Review, 26*, 90-102.
17. Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6,* 641-651.
18. Hagan, C. C., Woods, W., Johnson, S., Calder, A. J., Green, G. G. R., & Young, A. W. (2009). MEG demonstrates a supra-additive response to facial and vocal emotion in the right superior temporal sulcus. *Proceedings of the National Academy of Sciences, USA, 106*, 20010-20015.
19. Fruhholz, S., *et al.* (2015). Asymmetrical effects of unilateral right or left amygdala damage on auditory cortical processing of vocal emotions. *Proceedings of the National Academy of Sciences, USA. 112***,** 1583-1588.
20. Fruhholz, S., Trost, W., & Kotz, S.A. (2016). The sound of emotions: towards a unifying neural network perspective of affective sound processing. *Neuroscience and Biobehavioral Reviews, 68***,** 96-110.