Acta Psychologica 177 (2017) 54–68

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Facial biases on vocal perception and memory

Marilyn G. Boltz*

Haverford College, United States

ARTICLE INFO

Keywords:
Face-voice integration
Bimodal interactions
Bias
Social categorization
Vocal perception
Aging

ABSTRACT

Does a speaker's face influence the way their voice is heard and later remembered? This question was addressed through two experiments where in each, participants listened to middle-aged voices accompanied by faces that were either age-appropriate, younger or older than the voice or, as a control, no face at all. In Experiment 1, participants evaluated each voice on various acoustical dimensions and speaker characteristics. The results showed that facial displays influenced perception such that the same voice was heard differently depending on the age of the accompanying face. Experiment 2 further revealed that facial displays led to memory distortions that were age-congruent in nature. These findings illustrate that faces can activate certain social categories and preconceived stereotypes that then influence vocal and person perception in a corresponding fashion. Processes of face/voice integration are very similar to those of music/film, indicating that the two areas can mutually inform one another and perhaps, more generally, reflect a centralized mechanism of cross-sensory integration.

During everyday social interaction, we frequently encounter individuals for the first time and, because they are unfamiliar, form impressions of who they are and what they are like as a person. At this zero acquaintance level, such impressions are often based on facial and vocal qualities, simply because it is that information which is most readily and immediately available. The question addressed here is how the presence of both modalities may influence person perception and memory and, in particular, the potential effects of facial information on the vocal perception and remembering of a speaker's age. Such research not only has implications for impression formation but also more applied contexts such as ear and eyewitness testimony.

1. Behavioral influences of facial and vocal information

1.1. Unimodal presentations

Over the past several decades, a plethora of research has revealed that a remarkable amount of information can be reliably inferred about an individual simply by seeing their face or hearing their voice. In general, much of this research has been unimodal in nature in that participants are exposed to faces or voices alone while performing some sort of perceptual judgment task. The typical methodological strategy has been to present participants with a set of voices (speaking the same linguistic content) or faces belonging to individuals who are known to vary along a given dimension. The participants' task is to then rate or categorize each face (or voice) into its particular social group (e.g., male vs. female, African vs. European American) and if a high degree of accuracy is observed, conduct structural analyses to identify those features common among the set of experimental stimuli. As a converging operation, these particular facial or vocal features may later be manipulated within a presented set of experimental stimuli to determine if the same pattern of responses occurs.

Of the two modalities, the literature on facial perception is much more extensive, dating back to Darwin's (1872) book, *The Expression of the Emotions in Man and Animals*. The ability to recognize known individuals is necessary for social competency and familiar faces are, in fact, rapidly recognized and identified with a high degree of accuracy (e.g., Herzmann, Schweinberger, Sommer, & Jentzsch, 2004; Young, Hay, McWeeny, Flude, & Ellis, 1985). But in addition to personal identity, there are several other characteristics that can be accurately discerned and, as one might expect from an evolutionary perspective, many of these have survival and reproductive value. For example, a person's sex can be readily determined from bone structure and skin quality (Bruce et al., 1993; Russell & Sinha, 2007), while qualities such as cardiodal strain (George & Hole, 1995; Pittenger, Shaw, & Mark, 1979) and changes in skin pigmentation and elasticity (Burt & Perrett, 1995; Fink, Grammer, & Matts, 2006) provide cues for an individual's age. Similarly, other relevant characteristics such as emotional affect (Calvo & Lundqvist, 2008; Lederman, Kilgour, Kitada, Klatzky, & Hamilton, 2007), genetic fitness and health (Bulpitt, Markowe, & Shipley, 2001; Hwang, Atia, Nisenbaum, Pare, & Joordens, 2011), and ethnicity (Eberhardt, 2005; Hill, Bruce, & Akamatsu, 1995) offer highly useful information to others and these too are typically discerned with a high degree of accuracy. The dispositional traits of an individual are more difficult to detect but, nonetheless, there is evidence

* Corresponding author at: Department of Psychology, Haverford College, 370 Lancaster Avenue, Haverford, PA 19041, United States.
E-mail address: mboltz@haverford.edu.

http://dx.doi.org/10.1016/j.actpsy.2017.04.013
Received 19 December 2016; Received in revised form 17 April 2017; Accepted 27 April 2017
Available online 05 May 2017
0001-6918/

that the face conveys information about a person's level of intelligence (Zebrowitz, Hall, Murphy, & Rhodes, 2002; Zebrowitz & Rhodes, 2004), trustworthiness (DeBruine, 2005; Stirrat & Perrett, 2010), dominance (Carré, McCormick, & Mondloch, 2009; Weston, Friday, & Liò, 2007), and extroversion (Penton-Voak, Pound, Little, & Perrett, 2006).

Many of these same characteristics are also conveyed by the human voice, leading some to refer to it as the "auditory face" (Belin, Bestelmeyer, Latinus, & Watson, 2011). For example, research has shown that listeners are able to reliably determine a speaker's sex (Bachorowski & Owren, 1999; Smith & Patterson, 2005), particular identity (Papcun, Kreiman, & Davis, 1989; Schweinberger, Herholz, & Sommer, 1997), age (Torre & Barlow, 2009; Waller, Eriksson, & Sörqvist, 2015), ethnicity (Thomas, Lass, & Carpenter, 2010; Thomas & Reaser, 2004), and emotional state (Bachorowski, 1999; Juslin & Laukka, 2003; Scherer, 1986) from various pitch and temporal qualities of vocal information. In addition, other types of socially relevant information can be discerned that include an individual's relative body size (Fitch & Giedd, 1999; Smith & Patterson, 2005), sexual orientation (Gaudio, 1994; Munson, McDonald, DeBoe, & White, 2006; Tracy, Bainter, & Satariano, 2015), and geographical origin (Clopper & Pisoni, 2004; Garrett, Coupland, & Williams, 1999).

In sum, then, merely having access to a person's face or voice, with no additional interaction, can provide a wealth of information about that individual which, in turn, confers greater predictability to social situations and allows one to modify their own behavior appropriately.

This more traditional line of research reflects situations in which individuals have access to one modality alone, situations that do, in fact, occur in everyday life. There are occasions in which we form impressions of others based on a telephone exchange or an overheard conversation between two unseen individuals and, conversely, cases in which we are able to see the person but cannot hear their voice due to distance and/or background noise. However, it is perhaps more typical that we have joint access to both the face and voice when interacting with others. This, in turn, is advantageous in that redundant information between the facial and vocal channels as well as independent and non-overlapping information from the two modalities can provide a more reliable and diagnostic percept of an individual.

1.2. Bimodal presentations

Although the literature on bimodal presentations is more recent and therefore less extensive than the unimodal case, some research has nonetheless examined effects due to the joint presence of facial and vocal information. Much of this work has focused on speech perception (e.g., Campbell, 2008) but others have investigated effects on person identification (e.g., Ellis, Jones, & Mosdell, 1997) and the communication of emotional affect (Collingnon et al., 2008; de Gelder & Vroomen, 2000). In general, two different methodological strategies have been adopted in order to address two main questions of interest.

The first issue concerns the joint influence of congruent facial and vocal information relative to the face or voice alone. Does a bimodal vs. unimodal presentation facilitate or interfere with cognitive processing abilities? Although the latter has been observed in at least one study (Joassin, Maurage, Bruyer, Crommelinck, & Campanella, 2004), most of the evidence points to a facilitating effect. For example, in speech perception, the availability of a speaker's face and accompanying lip, head, and facial movements aids phoneme identification, the recognition of words within a sentence, and segmentation of the speech stream (e.g., Massaro & Cohen, 1999; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). As noted by Yehia, Rubin, and Vatikiotis-Bateson (1998), this phenomenon is not particularly surprising given that vocal tract movements influence both facial movements as well as the produced speech sounds themselves. In the context of person identification, the presence vs. absence of an individual's face not only aids the learning and remembering of unfamiliar voices (Sheffert & Olson, 2004), but increases the speed at which a familiar voice is identified as such (Ellis et al., 1997). Similarly, the speed at which participants are able to accurately categorize the gender (e.g., Joassin, Maurage, & Campanella, 2011; Latinus, VanRullen, & Taylor, 2010) or various emotional expressions (e.g., Collingnon et al., 2008; de Gelder & Vroomen, 2000) of an individual is significantly faster in the presence of bimodal vs. unimodal presentations.

On a theoretical level, these types of facilitating effects have been addressed in a neurocognitive model of vocal perception developed by Belin, Fecteau, and Bedard (2004) that parallels that of Bruce and Young (1986) for facial perception and, most importantly, integrates the two sources of information. The basic idea is that an initial low-level analysis occurs for each modality followed by a structural analysis in which three primary types of information are extracted for further processing, namely, speech, affective, and identity information – all of which contribute to person perception. Although speech, affect, and identity are assumed to be processed independently of one another within a given modality, they also interact both within and between modalities at a neural level. In particular, neuroimaging data has revealed temporal voice areas (TVAs) along the bilateral superior temporal sulcus that respond more strongly to vocal vs. nonvocal sounds (e.g., Belin, Zatorre, Lafaille, Ahad, & Pike, 2000) and display connectivity to the fusiform face area (e.g., von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). Several other neural regions have been identified in the integration of facial and vocal information and as described in a review paper by Campanella and Belin (2007), involve a network of both cortical and subcortical regions.

In addition to bimodal vs. unimodal presentations, a second but less frequently used methodological strategy has been to present bimodal facial and vocal information that is congruent vs. incongruent with one another. The aim here is to examine whether one modality exerts a greater effect than the other and thereby dominates for a given type of behavior. Perhaps the two most well-known examples of this approach are the ventriloquism effect in which audiovisual discrepancies in spatial location are resolved through visual dominance wherein participants "hear" sound emanating from the moving lips of a dummy (e.g., Bertelson & Radeau, 1981). Similarly, in the McGurk effect (McGurk & MacDonald, 1976), participants who hear a given phoneme (e.g., "ba") accompanied by a video of a speaker producing a different phoneme (e.g., "ga") resolve this sensory conflict through a merging of the two modalities (i.e., hearing "da"). Such biases are not specific to the realm of speech but have also been found in the domain of person perception. For example, in a study by Latinus et al. (2010), participants were presented with faces and voices that were gender congruent or incongruent and, while attending to one modality alone, asked to categorize each stimulus as male or female. The results showed that the face dominated the voice in gender perception: incongruent faces disrupted the gender categorization of voices but not vice versa. Similar effects have been observed for emotional affect wherein the affective expression of faces alters vocal perception, even when instructed to ignore the face (de Gelder, Pourtois, & Weiskrantz, 2002; Vroomen, Driver, & deGelder, 2001).

More recently, through a clever research methodology, some have examined the dynamic time course in which congruent vs. incongruent face-voice information is integrated and resolved (e.g., Freeman & Ambady, 2011). By tracking the hand movements guiding a computer mouse toward two categorical responses (e.g., male, female) displayed on-screen, one finds that the integration of the two modalities is an ongoing and continuous process that occurs quite quickly after face/voice onset (i.e., 250 ms, on average) and, as one might expect, more quickly for congruent than incongruent pairings. As noted by these authors, this type of methodology may be useful for investigating populations in which the ability to integrate face-voice information is impaired such as those of schizophrenics and alcoholics.

In sum, then, the research on unimodal presentations has been very useful in identifying those personal characteristics that can be accurately and reliably inferred from the face and voice as well as those structural properties that afford such information. The more recent trend of examining bimodal interactions enhances the ecological