

**Kognitive Verarbeitung von Stimminformation**

**(Cognitive mechanisms of voice processing)**

Inaugural-Dissertation

zur Erlangung der Doktorwürde der

Fakultät für Humanwissenschaften

der

Julius-Maximilians-Universität Würzburg

Vorgelegt von

Sujata Maya Huestegge

aus Aachen

Würzburg

2019



Erstgutachterin: Professor Dr. Gerhild Nieding

Zweitgutachterin: Professor Dr. Wafaa Shehata-Dieler

Tag des Kolloquiums: 25.07.2019

## **Vorbemerkung**

Die vorliegende Dissertation mit dem Thema „Kognitive Verarbeitung von Stimminformation“ wird hiermit in teilkumulativer Form eingereicht. Neben dem folgenden Mantelteil, der den theoretischen Rahmen und Zusammenhang der erfolgten empirisch-experimentellen Studien verdeutlicht, umfasst diese Arbeit drei Artikel. Zwei dieser Artikel (bei denen ich als Hauptautorin fungiere) befinden sich bei international anerkannten Zeitschriften mit *peer review*-Verfahren bereits „im Druck“:

**Huestegge, S. M., & Raettig, T.** (*in press*). Crossing gender borders: Bidirectional dynamic interaction between face-based and voice-based gender categorization. *Journal of Voice*. (akzeptiert am 25.09.2018)

DOI der *online first*-Publikation: <https://doi.org/10.1016/j.jvoice.2018.09.020>

**Huestegge, S. M., Raettig, T., & Huestegge, L.** (*in press*). Are face-incongruent voices harder to process? Effects of face-voice gender incongruency on basic cognitive information processing. *Experimental Psychology*. (akzeptiert am 10.01.2019)

Ein weiterer Artikel in Alleinautorschaft ist bei einer international anerkannten Zeitschrift mit *peer review*-Verfahren eingereicht worden und befindet sich im Begutachtungsprozess:

**Huestegge, S. M.** (*submitted*). Matching unfamiliar voices to static and dynamic faces: No evidence for a dynamic face advantage in a simultaneous presentation paradigm. *Frontiers: Psychology*. (eingereicht am: 15.02.2019)

Alle Studien dieser teilkumulativen Dissertation werden in der hier vorliegenden Form mit Erlaubnis der jeweils zuständigen Verlage verwendet.

## Inhaltsverzeichnis (Contents)

<b>0.1 Zusammenfassung</b>	<b>5</b>
<b>0.2 Summary</b>	<b>7</b>
<b>1. General introduction: Cognitive processing of voice information</b>	<b>9</b>
<b>2. Basic views on cognition: Modular, abstract vs. interactive, embodied processing</b>	<b>10</b>
<b>2.1 Modular, abstract information processing</b>	<b>10</b>
2.1.1 Modular cognitive organization	10
2.1.2 Abstract information processing	12
<b>2.2 Interactive, embodied processing</b>	<b>13</b>
2.2.1 Strongly interactive organization in cognition	13
2.2.2 Embodied and situated cognition	14
<b>3. The case of voice (and face) processing</b>	<b>16</b>
<b>3.1 Modular, abstract organization of face and voice information processing</b>	<b>16</b>
<b>3.2 Evidence for strongly interactive, embodied organization of face and voice information processing</b>	<b>17</b>
3.2.1 Face-voice speech integration	17
3.2.2 Face-voice gender integration	18
3.2.3 Face-voice integration for other features	22
3.2.4 Evidence from face-voice matching	22
<b>4. Theories of voice (and face) processing and implications of the present studies</b>	<b>26</b>
<b>4.1. Introductory remarks: What's in a voice? General aspects of cognitive voice processing</b>	<b>26</b>
<b>4.2 Voice (and face) processing theories</b>	<b>27</b>
<b>4.3 Situating the present research questions in the models</b>	<b>30</b>
4.3.1 Study 1 (Huestegge & Raettig, in press): Gender-related face-voice processing interaction	31
4.3.2 Study 2 (Huestegge et al., in press): Impact of gender-congruency on basic, speech-based information processing	33
4.3.3 Study 3 (Huestegge, subm.): Matching faces and voices	34
<b>5. Results of the current studies and theoretical implications</b>	<b>36</b>
<b>5.1 Results &amp; implications of Study 1 (Huestegge &amp; Raettig, in press)</b>	<b>36</b>
<b>5.2 Results &amp; implications of Study 2 (Huestegge et al., in press)</b>	<b>39</b>
<b>5.3 Results &amp; implications of Study 3 (Huestegge, subm.)</b>	<b>42</b>
<b>6. Final discussion: Novel theoretical insights</b>	<b>44</b>
<b>7. References</b>	<b>49</b>

## 0.1 Zusammenfassung

Die vorliegende Dissertation thematisiert die kognitive Verarbeitung von Stimminformation. Basierend auf allgemeinen theoretischen Vorstellungen zu mentalen Prozessen wird zunächst unterschieden in modulare, abstrakte Informationsverarbeitungsansätze und interaktive, verkörperte Vorstellungen kognitiver Prozesse. Diese allgemeinen Vorstellungen werden dann am Beispiel der Verarbeitung von Stimminformation im Kontext der parallel dazu ablaufenden Gesichtsverarbeitung konkretisiert. Es geht also u.a. darum, inwiefern kognitive Stimmverarbeitung unbeeinflusst von der gleichzeitigen Verarbeitung von visueller Personeninformation ablaufen kann (und umgekehrt). In Studie 1 (Huestegge & Raettig, in press) werden Probanden audiovisuelle Stimuli dargeboten, bei denen Gesichter Ziffern aussprechen. Manipuliert wird die Geschlechtskongruenz der Stimuli: Es gibt männliche und weibliche Gesichter, die je entweder mit einer männlichen oder weiblichen Stimme synchronisiert wurden. Probanden sollen entweder nur auf die Stimme oder nur auf das visuelle Gesicht achten und jeweils das Geschlecht per Tastendruck kategorisieren. Dabei stellte sich heraus, dass es für die Kategorisierungsleistung eine Rolle spielt, ob es sich um geschlechts-kongruente oder -inkongruente Stimuli handelt: Letztere wurden langsamer bzw. mit höherer Fehleranfälligkeit kategorisiert, was für eine starke cross-modale Interaktion der zugrundeliegenden visuellen und akustischen Verarbeitungsrouten spricht. Dabei wirkte sich inkongruente visuelle Information stärker auf die Stimmbeurteilung aus als inkongruente Stimminformation auf die visuelle Beurteilung, was auf eine Dominanz visueller gegenüber akustischer Informationsverarbeitung hindeutet. Unter starker kognitiver Belastung konnte ebenfalls ein Kongruenzeffekt nachgewiesen werden. In Studie 2 (Huestegge, Raettig, & Huestegge, in press) wurde dasselbe Stimulusmaterial verwendet, aber kategorisiert werden sollten nun die gesprochenen Ziffern (z.B. in gerade/ungerade oder größer/kleiner 5). Damit ist in der Instruktion die Aufmerksamkeit

von der Geschlechtsdimension weggelenkt. Dennoch fanden sich auch hier Geschlechtskongruenzeffekte auf die Ziffernkategorisierung, was für eine relativ automatische Verarbeitung von cross-modaler Geschlechtsinformation spricht, die sich dann auch auf die Sprachverarbeitung auswirken kann. In Studie 3 (Huestegge, subm.) wurde die Fähigkeit von Probanden untersucht, von einer Stimme auf das zugehörige (statisch oder dynamisch) dargebotene Gesicht zu schließen. Dies gelang den Probanden in überzufälliger Weise. Weiterhin konnte keine Evidenz dafür gefunden werden, dass bewegte (dynamische) Gesichter besser den Stimmen zugeordnet werden konnten als statische Gesichter. Die Ergebnisse sprechen dafür, dass gemeinsame Quellinformation sich sowohl auf Stimme wie Gesichtsmerkmale auswirkt, und dass implizites Wissen hierüber von den Probanden genutzt wird, um Stimmen Gesichtern zuzuordnen. Insgesamt konnten die Ergebnisse der drei Studien (Huestegge, subm.; Huestegge & Raettig, in press; Huestegge et al., in press) dazu beitragen, bestehende Theorien der Stimm- und Gesichterverarbeitung entscheidend weiterzuentwickeln. Die Ergebnisse sind allgemein eher im Einklang mit einer stark interaktiven, verkörperten Sicht auf kognitive Prozesse, weniger mit einer modular-abstrakten Informationsverarbeitungsperspektive.

## 0.2 Summary

The present thesis addresses cognitive processing of voice information. Based on general theoretical concepts regarding mental processes it will differentiate between modular, abstract information processing approaches to cognition and interactive, embodied ideas of mental processing. These general concepts will then be transferred to the context of processing voice-related information in the context of parallel face-related processing streams. One central issue here is whether and to what extent cognitive voice processing can occur independently, that is, encapsulated from the simultaneous processing of visual person-related information (and vice versa). In Study 1 (Huestegge & Raettig, in press), participants are presented with audio-visual stimuli displaying faces uttering digits. Audiovisual gender congruency was manipulated: There were male and female faces, each uttering digits with either a male or female voice (all stimuli were AV- synchronized). Participants were asked to categorize the gender of either the face or the voice by pressing one of two keys in each trial. A central result was that audio-visual gender congruency affected performance: Incongruent stimuli were categorized slower and more error-prone, suggesting a strong cross-modal interaction of the underlying visual and auditory processing routes. Additionally, the effect of incongruent visual information on auditory classification was stronger than the effect of incongruent auditory information on visual categorization, suggesting visual dominance over auditory processing in the context of gender classification. A gender congruency effect was also present under high cognitive load. Study 2 (Huestegge, Raettig, & Huestegge, in press) utilized the same (gender-congruent and -incongruent) stimuli, but different tasks for the participants, namely categorizing the spoken digits (into odd/even or smaller/larger than 5). This should effectively direct attention away from gender information, which was no longer task-relevant. Nevertheless, congruency effects were still observed in this study. This suggests a relatively automatic processing of cross-modal gender information, which

eventually affects basic speech-based information processing. Study 3 (Huestegge, *subm.*) focused on the ability of participants to match unfamiliar voices to (either static or dynamic) faces. One result was that participants were indeed able to match voices to faces. Moreover, there was no evidence for any performance increase when dynamic (vs. mere static) faces had to be matched to concurrent voices. The results support the idea that common person-related source information affects both vocal and facial features, and implicit corresponding knowledge appears to be used by participants to successfully complete face-voice matching. Taken together, the three studies (Huestegge, *subm.*; Huestegge & Raettig, *in press*; Huestegge et al., *in press*) provided information to further develop current theories of voice processing (in the context of face processing). On a general level, the results of all three studies are not in line with an abstract, modular view of cognition, but rather lend further support to interactive, embodied accounts of mental processing.



## **1. General introduction: Cognitive processing of voice information**

The present work operates on three levels. First, on the most general theoretical level, it discusses several fundamental concepts of human cognition, addressing the basic questions of whether our cognitive system is designed in a highly modular, or, alternatively, in a strongly interactive way with strong crosstalk between streams of information processing. Furthermore, I address the perennial issue of whether central cognitive processes are based on abstract representations of task-relevant information (“information processing” approach to cognition), or whether cognition is instead essentially embodied and situated. Second, on a more domain-specific theoretical level, these general questions are narrowed down to a particular research field, namely voice (and face) processing. Since basic cognitive theories of voice processing are inextricably linked to (and derived from) face processing theories, and since both face and voice processing are usually devoted to very similar types of information about persons, theories of voice processing will always be treated in conjunction with face processing theories. Third, on an empirical level, I will describe several experiments (Studies 1-3) that address both the general theoretical issues referred to above as well as more specific aspects derived from current theories of face-voice processing. In this summary, I aim to show how the several experiments described in the manuscripts address specific open theoretical issues in the context of current face-voice processing theories. Finally, they are suited to augment and advance these theories, as will be shown in the final chapter of this theoretical summary.

## 2. Basic views on cognition: Modular, abstract vs. interactive, embodied processing

When listening to the voices of others while communicating, we usually do not process auditory signals alone, but instead integrate pieces of information from different input channels at the same time, including information from the visual processing channel.

Regarding the basic underlying cognitive architecture of multi-channel processing, two views can be distinguished: a modular view and a strongly interactive view of cognition.

Historically, the modular view is also loosely tied to the idea of rather abstract information processing, as opposed to a situated, embodied view of cognition. These ideas will be outlined in the following, before focusing on the specific case face and voice processing.

### 2.1 Modular, abstract information processing

In this section, the notion of modular, abstract information processing will briefly be outlined.

**2.1.1 Modular cognitive organization.** According to a heavily *modular* account of cognition (Fodor, 1983), it should be possible to process voice information largely independently from processing other streams of information, especially when different input channels are involved. In the early days of cognitive psychology, Broadbent (1958) proposed that people should be able to easily filter their attention based on (e.g., visual and auditory) input channels or modules. A signature feature of modularity is encapsulation (Fodor, 1983), that is, the property of modules to be immune to interference in terms of influence from other ongoing processes.

The claim that the processing of relevant stimulus information for a specific module should not be subject to interference from irrelevant stimulus information from other modules is also plausible based on findings from a long tradition of research on fundamental attention mechanisms. Specifically, research on visual attention (using the visual search paradigm) has

established the general rule that input processing devoted to a single defined feature (feature search) does not call for attentional resources, since the relevant feature should “pop out” of the distracting, task-irrelevant information. In contrast, resource-consuming attentional demands are present when a current target stimulus is defined in terms of a conjunction of features (conjunction search, e.g., when a participant is asked to only respond to a certain digit when this digit is uttered by a male face, so that a conjunction of a visual feature and an auditory feature must be considered at the same time in order to respond accordingly). However, when only one feature (e.g., the auditory digit information) is task-relevant, corresponding information processing should not require attentional demands and thus rather lead to a conceptual equivalent of a “pop out” effect (Treisman & Gelade, 1980; Wolfe, 1994) regarding the relevant digit information.

Corresponding formal theoretical frameworks (computational approaches) have also been proposed in other, even more general theories of attention, where selective attention is reflected in processing prioritization parameters, enabling subjects to process relevant and ignore irrelevant information (e.g., see the Theory of Visual Attention [TVA] by Bundesen, 1990). Taken together, these lines of theoretical reasoning predict that humans are usually characterized by the ability to perfectly focus on task-relevant information, especially when only one feature is task-relevant (see Study 2 of this thesis: Huestegge, Raettig, & Huestegge, in press) and when the task-irrelevant information is strongly separated (e.g., by being processed by another sensory system) from the relevant source of information. As mentioned above, the idea that attention should be able to filter exceptionally well by input modality was already postulated by Broadbent (1958). Probably, this might be a reason why many well-known basic interference phenomena (Eriksen flanker effect, Stroop effect, Simon effect, see Kornblum, Hasbrouck, & Osman, 1990, for an overview and details) are typically characterized by the fact that interference occurs within the *same* (e.g., visual) processing channel, and the irrelevant information is usually strongly associated with the relevant

information by sharing a conceptual dimension (space, color etc.), but with opposite features (responding to a left target with the right hand, saying “blue” in response to the word red printed in blue etc.). Thus, dimensional overlap can be considered a potential threat to perfect filtering abilities. Nevertheless, cross-modal attention studies demonstrating interference between different input modalities also emerged, which will be outlined in more detail in the corresponding chapter on interactive processing (see Section 2.2.1).

Typical models of modular cognition are box-and-arrow models, in which the boxes typically represent modules, whose output is transferred (as indicated by arrows) to other modules. One example for such a functional processing model is Bruce and Young’s (1986) face processing model described in Section 4.2.

**2.1.2 Abstract information processing.** Historically, a rather modular view of cognition (also involving modules such as “working memory” or “long-term memory”, which represent metaphors borrowed from the emerging research field of computer sciences) was at its prime at the same time when information processing was considered to be basically abstract in nature (similar to bits in digital processing computers, see Marx & Cronan-Hillix, 1987). A central claim of the corresponding field of information processing psychology is that particular input and output modalities are not relevant for central, cognitive information processing: All that counts is the task-relevant information, irrespective of which channel it came from, and irrespective of which motor system has to execute selected responses. This idea is based on a longstanding tradition in psychology that emerged with the “cognitive turn” around 1960, when psychological theory has turned away from behaviorism and towards a new science of the mind by utilizing information processing in the computer as a general metaphor for cognitive “computations” in the mind (Marx & Cronan-Hillix, 1987). Based on this metaphor, it has been assumed that cognition emerges based on an abstract central processing unit (CPU, offering central computation resources or processing capacity), similar

to a computer CPU that only processes binary, abstract units of information (bits) without any resemblance of the stimuli that triggered these bits of information, and without any resemblance of what these bits eventually trigger (e.g., letters on the screen in the case of text writing software).

Especially Fodor (1975) has promoted the idea that central mental processes are coded in an abstract cognitive, formalized language (“language of thought”, or “mentalese”, somewhat similar to binary processing language in information technology systems). A central prediction here is that task-irrelevant stimulus information should not be considered (especially when assuming that filtering relevant information is comparatively easy), since only relevant information should be transformed into abstract cognitive codes. Of course, such abstract information processing theories would also predict that the *relation* between two task-irrelevant stimulus dimensions (i.e., their (in)congruency) – in the absence of dimensional overlap – should also not affect the processing of task-relevant information. This idea will be explicitly tested in Study 2 (Huestegge et al., in press), and it is especially plausible in the light of many typical cognitive conflict tasks (such as Stroop tasks, Stroop, 1935, or Flanker tasks, Eriksen & Eriksen, 1974), which always involve dimensional overlap between task-relevant and task-irrelevant stimulus dimensions (making filtering relevant information more difficult, see Kornblum, Hasbrouck, & Osman, 1990).

## **2.2 Interactive, embodied processing**

In this section, the notion of interactive and embodied processing will be introduced.

**2.2.1 Strongly interactive organization in cognition.** However, there is also some evidence for a highly *interactive* account of cognition, especially with respect to two probably most important input channels of human cognition: visual and auditory processing. Typically, this claim is supported by well-known phenomena of cross-talk between both processing channels. For example, cross-modal attention research and studies on multisensory integration

have shown that when an auditory stimulus is accompanied by a visual stimulus, processing of the former can be strongly attenuated or even completely neglected (Colavita visual dominance effect; Colavita, 1974; see Spence, 2009). In addition, the famous ventriloquism effect shows that localization of sounds is biased towards the location of visual objects that are presented at the same time (e.g., Bertelson & Aschersleben, 1998; Warren, Welch, & McCarthy, 1981; also see Shams, Kamitani, & Shimojo, 2002, for a corresponding reversed effect of auditory information on visual processing). Further effects of cross-modal interference for face and voice processing will be presented in the following chapters.

A typical class of models assuming strongly interactive cognition are parallel distributed processing (PDP) models (sometimes also referred to as connectionist models). These models operate with codes (nodes which represent any type of information), and these codes are connected with variable connection strength, similar to actual neural networks (which actually served as a metaphor for these models). Such PDP models can easily explain both priming (i.e., facilitatory effects) and interference effects (i.e., adverse effects) by assuming bidirectional spreading of activation (or inhibition) between nodes in the network (e.g., McClelland & Rumelhart, 1981).

**2.2.2 Embodied and situated cognition.** The aforementioned accounts of cognition have in common that they involve an abstract view of information processing, which is conceptualized in terms of task-specific processing codes and potential dimensional overlap of code sets. However, there is also an alternative theoretical view that gained traction in recent years, namely an *embodied* cognition perspective (e.g., Barsalou, 2008, Barsalou, Simmons, Barbey, & Wilson, 2003). A central claim of this perspective is that any mental representations are grounded in (or coded in terms of) their sensory and/or motor origins, and that these mental representations are embedded in their respective situational context (situated cognition, see Wilson, 2002). Thus, unlike abstract information processing theories of

cognition (see Section 2.1.2), these theories assume that (central cognitive) information processing is never abstract, but cognitive processes always take their sensory/motor origins into account. This particular viewpoint will be further specified in the following in the context of voice (and face) processing models.

### **3. The case of voice (and face) processing**

The general concepts developed in the previous chapter can also be applied to the more specific field of the present thesis, that is, the field of voice and face processing.

#### **3.1 Modular, abstract organization of face and voice information processing**

Considerable evidence has been presented for the assumption of a modular cognitive organization of person-related processing, in particular with respect to face and voice processing. Some of this evidence comes from brain research: Functional neuroanatomy indicates that separate brain networks are involved in face processing (e.g., fusiform face area) and voice processing (e.g., temporal voice area, see Yovel & Belin, 2013, for a review of corresponding evidence). Further evidence is based on neuropsychological patients. Specifically, there are patients who show a strong impairment of face-based visual processing (a disorder referred to as prosopagnosia), while their general object recognition abilities as well as voice recognition abilities are fully intact (e.g., Barton, 2008). Moreover, there is a neuropsychological disorder named “phonagnosia”. These patients cannot recognize other individuals when hearing their voice. However, their face and name processing abilities are typically unimpaired (e.g., Garrido et al., 2009; Hailstone, Crutch, Vestergaard, Patterson, & Warrena, 2010; Herald, Xu, Biederman, Amir, & Shilowich, 2014; Van Lancker, Cummings, Kreiman, & Dobkin, 1988).

In response to this evidence for modularity, the most influential theory of face processing has been introduced by Bruce and Young (1986), a theory that will be outlined in more detail later. This theory was developed further so that it also accounts for voice processing (e.g., Stevenage, Hugill, & Lewis, 2012). This line of theory assumes parallel, but relatively separate pathways for processing faces and voices to account for the modularity found in the brain and with respect to patients (see above). A central feature in these theories is that face and voice processing pathways can be *selectively* activated (Yovel & Belin, 2013),



that is, in a modular manner. Nevertheless, integrative person perception is still possible because all processing pathways are connected to central cognitive processes (Bruce & Young, 2012).

In line with these modular face and voice processing theories, early research on face and voice processing indicated that both types of processing can indeed occur relatively independently. Specifically, corresponding studies reported only very limited evidence for cross-modal priming (i.e., priming of voice processing by using face primes and vice versa), which could only be observed in few, very special conditions (Ellis, Jones, & Mosdell, 1997).

### **3.2 Evidence for strongly interactive, embodied organization of face and voice information processing**

In contrast to the modular accounts referred to above, there is also evidence for an alternative view that rather emphasizes interactions between several processing domains. Some of these ideas will be outlined in the following.

**3.2.1 Face-voice speech integration.** The probably most influential effect known from the literature of face-voice interaction is related to the processing of speech information, namely the McGurk effect. In the seventies of the last century, McGurk and MacDonald (1976) demonstrated that visual (facial) information can strongly affect auditory speech perception. In their study, they included stimulus material which shows a woman's face while uttering the syllables /ba-ba/, /ga-ga/, /pa-pa/, or /ka-ka/. These stimuli were AV-synchronized in a systematic manner, utilizing four combinations overall: a voice uttering /ba/ combined with visual face/lip movements uttering /ga/; a voice uttering /ga/ combined with lip movements for /ba/, a voice uttering /pa/ combined with lip movements for /ka/, and finally a voice uttering /ka/ combined with lip movements for /pa/. Participants (which consisted of adults, school children, and pre-school children) were simply instructed to verbally repeat what they heard from these AV-synchronized models. The most important result was that in

many instances (e.g., in 98% of responses from the adult group), the vocal utterance /ba/ combined with /ga/-lip movements resulted in a /da/-response from the participants. Thus, it appears that subjects tended to perceive a fused percept of the visual and auditory information, which can be interpreted as evidence for cross-modal integrative processing. An attenuated (but still impressive) effect was reported in the condition combining the auditory utterance /pa/ with lip movements from /ka/, where participants responded with the voiceless /ta/ in 81% of cases (adult group). It is interesting to note that these two conditions are similar in that they involved a bilabial plosive in the auditory information stream that was not accompanied by lip closing in the visual information stream. This combination, which does not occur in natural life, might have triggered perceptual disambiguation processes in the participants, which eventually may have resulted in the fused percept. Taken together, the McGurk effect is a famous instance of perceptual AV-integration, which involves the creation of an illusory perception of speech that is actually not present in either the auditory or visual domain.

Interestingly, a follow-up study by Green, Kuhl, Meltzoff, and Stevens (1991) showed that a McGurk effect is still present even when the stimulus material suggests that the voice and the face do not belong to the same person. Specifically, they combined female voices with male faces (and vice versa). Thus, it is not important for the effect to occur whether the participants believe that both streams of processing belong to the same person. This means that these cross-modal integration processes are relatively automatic and cannot be affected by higher cognitive processes such as beliefs about a person.

**3.2.2 Face-voice gender integration.** Apart from the McGurk effect that is related to speech processing, audio-visual integration phenomena were additionally observed in the context of gender information processing. One study (Joassin, Maurage, & Campanella, 2011) demonstrated that gender classification based on visual information can be completed

significantly faster when auditory (voice) information is presented in addition to the visual information (compared to presenting visual face information alone). Additionally, they found that mere auditory presentation (i.e., without visual information) was associated with the slowest response times.

Despite these relatively rare examples of *voice*-based gender classification studies in the context of additional visual face information, many experiments on gender classification are concerned with *face*-based gender perception in the context of additional auditory information, probably due to the fact that more researchers were coming from the field of face (compared to voice) processing (see also Yovel & Belin, 2013). One example is a study by Smith, Grabowecky, and Suzuki (2007), who reported that androgynous faces (i.e., those which are ambiguous with respect to gender) were more often classified as being male/female when these faces were shown in conjunction with tones of low/high frequency, respectively. An explanation by the authors for this effect was that the low/high tones are reminiscent of the male/female fundamental speaking frequency range, thereby priming corresponding gender categorization responses. Another study (Masuda, Tsujii, & Watanabe, 2005) showed that gender-incongruent voices modulated the N170 event-related potential component (which is assumed to be correlated with face processing) relative to gender-congruent voices. The authors conclude that this modulation might indicate a disruption of face-based gender processing. Finally, there is a further study which used morphed, gender-ambiguous faces (Freeman & Ambady, 2011). These faces were presented along with either original male/female voices or with morphed voices (these morphed voices had altered formants, thereby becoming gender-atypical). Similar to the Smith et al. (2007) results (see above), they showed that the particular voice information biased gender classification in that gender-atypical voices slowed down face-based gender categorization.

Interestingly, there is a research gap regarding effects of visual face-based gender information on voice-based gender classification performance. One previous study, however, was conducted in the context of a closely related issue. It addressed potential effects of visually presented gender on singing voice type categorization (specifically: bass, baritone, alto, soprano, Peynircioglu, Brent, Tatz, & Wyatt, 2017). In their experiment, they utilized short (1s) sung voice snippets that were sung by either male or female singers at constant pitch (g3). In one condition, voice type classification was only based on auditory information. In this condition, about one third of the subjects were unable to categorize the acoustic stimuli accurately, which indicates that (due to the constant, high pitch) these stimuli were quite ambiguous regarding voice type. In other conditions, these auditory tracks were synchronized with a (visual) video of a male or female singer. The data showed that categorization of voice type was shifted in the direction of the visually presented gender: For example, a male voice was more often categorized as soprano/alto when accompanied with a visual presentation of a female singer.

However, these previous studies may have a limited generalizability. Note that most of the stimuli used in such studies involved gender-ambiguous material, thereby creating a somewhat artificial situation. Specifically, it is possible that only when confronted with ambiguous stimuli subjects direct their attention towards any additional information (here: in the other input modality) that might support them to come up with a more informed, less ambiguous decision. It is therefore perfectly possible that interference effects (indexing integration processes) are present when ambiguous information is presented, but that in the absence of ambiguous information (which is the standard case in everyday life) modular, independent processing is perfectly possible. Thus, to advance the claim that audio-visual integration is a standard processing mode in person perception (also in everyday life), corresponding studies with unambiguous stimulus material are needed. Additionally, such studies should ideally look at both directions of influence (face-on-voice effects and vice

versa). One notable study that meets these requirements was conducted by Latinus, VanRullen, and Taylor (2010), who indeed looked at both directions of influence and also used non-ambiguous stimulus material. In particular, they utilized static photos of faces that were displayed in addition to either gender-congruent or gender-incongruent male/female voices. These voices uttered short French words. Their study design involved separate blocks of trials for either face or voice categorization demands. A central result of this study was that interference indeed worked in both directions. Additionally, visual-on-auditory interference effects were stronger than auditory-on-visual interference effects. This is in line with previous reports of visual dominance over auditory processing (e.g., Colavita, 1974).

Despite its pioneering character, this study by Latinus et al. (2010) also has some disadvantages. Most importantly, the interesting Colavita-like asymmetry effect is confounded with a modality-related difference in presentation conditions: There was no movement of the (static) faces that corresponded to the heard voice, which is inherently dynamic. Such a phenomenon is highly artificial in that it is not encountered in daily life situations, and – as a consequence – likely has directed more attention to the static face. In turn, this increased attention to the face might have yielded the reported visual dominance effect.

As a consequence, it appears that an ideal study would be similar to that of Latinus et al. (2010), but involve naturally moving visual faces instead of static faces. This would then for the first time enable a bi-directional study of actual face-voice *integration* processes, because only a simultaneous, dynamic speech presentation in both the visual and auditory processing channel will result in a more natural situation that is necessary for the study of real dynamic integration processes as envisioned, among others, by McGurk and MacDonald (1976). Such a study will be presented as Study 1 (Huestegge & Raettig, in press) in the present thesis.

**3.2.3 Face-voice integration for other features.** Apart from speech- and gender-related information, evidence for face-voice integration (or audio-visual integration) in terms of congruency effects has also been presented for other types of information. For example, congruency effects were reported for emotional (affective) information (de Gelder & Vroomen, 2000; Hagan et al., 2009; Hietanen, Leppänen, Illi, & Surakka, 2004; Purtois et al., 2005), age-related information (Boltz, 2017), and information related to a person's identity (in terms of familiarity, i.e., there was evidence for modulated processing of familiar speakers presented with identity-corresponding vs. -noncorresponding speaking faces, Schweinberger, Kloth, & Robertson, 2011). Finally, by using a visual search for faces task, it has been shown that the presence of a corresponding voice speeds up visual search for a corresponding face (Zweig, Suzuki, & Grabowecky, 2015). However, since these types of information are not directly relevant for the present thesis, I will not present these studies in more detail (but note that in the frameworks of voice processing in the context of person perception these processing dimensions are also incorporated, see Section 6).

**3.2.4 Evidence from face-voice matching.** The studies outlined in the following are based on the central assumption that faces and voices share common, person-related source identity information, thereby enabling above-chance abilities to match previously unknown faces and voices. A pioneering attempt to develop a corresponding paradigm was introduced by Lachs (1999), who described a sequential face-voice matching task based on classic match-to-sample paradigms in psychology. This procedure was afterwards utilized in many other studies looking at performance for matching either static or dynamic faces to corresponding voice samples (e.g., Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004 a,b; Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007). A typical corresponding paradigm involves several sequential steps in each trial: First, a reference stimulus in one modality (e.g., a visual face stimulus) is shown. Then, a sequence of two comparison stimuli in the other modality are presented (e.g., two alternative voices), one after

another. After the presentation of these two alternatives, a final decision phase requires the participant to indicate with of the two comparison stimuli corresponds to the reference stimulus. This is usually designed as a two-alternative forced choice task, that is, one of the two alternatives is always correct. In this paradigm, it is possible to manipulate the order of modalities (e.g., presenting an auditory or visual reference stimulus prior to the comparison stimuli in the other modality) while keeping the overall procedure constant. It is also possible to use static or dynamic visual face information in this task.

Most studies using this (or a comparable) task reported performance at chance level when static instead of dynamic visual faces were presented, with the exception of a study by Mavica and Barenholtz (Exp. 2). However, most studies agree that performance was clearly above chance when using dynamic visual faces. These findings were interpreted by assuming that dynamic face information is a *conditio sine qua non* for finding above-chance-level matching performance (e.g., Kamachi et al., 2003), probably because the relevant common source information is dynamic in nature (e.g., speech tempo, or transient information etc.).

However, other studies offered a different potential explanation for the observations of chance-level performance in the context of static face stimuli. In particular, Smith, Dunn, Bagulay, and Stacey (2016a) speculated that the four phases involved in the classic paradigm (reference stimulus, comparison stimulus 1, comparison stimulus 2, response) could have resulted in a strong load for working memory. This in turn might particularly impact on conditions involving static face presentation, which might represent a harder memory demand in the first place. In line with this assumption, findings from research on memory processes indicated worse general memory performance for static faces than for dynamically presented faces (Christie & Bruce, 1998; Knappmeyer, Thornton, & Bülthoff, 2003; Lander & Chuang, 2005), a finding that has been explained by assuming that dynamically presented faces contain more retrieval cues indexing a certain person (O'Toole, Roark, & Abdi, 2002). Of

course, this reasoning would predict that face-voice matching paradigms that lower working memory demands should reveal above-chance matching performance for static faces, too.

This prediction was indeed corroborated by studies which involve a simultaneous display of the visual comparison stimuli (note that this procedure is not feasible for auditory comparison stimuli, since simultaneous presentation of two audio tracks will severely hamper the processing of individual voice features): Here, above-chance matching performance was found, and this result was replicated across different research groups (Krauss, Freyberg, & Morsella 2002; Mavica & Barenholtz, 2013; Smith et al., 2016a, Exp. 3). Note that some of these studies had specific characteristics: For example, Krauss et al. (2002) displayed not only faces, but complete bodies of model persons. Mavica and Barenholtz (2013, Experiment 1) introduced a special design where not only the two comparison stimuli, but also the reference stimulus were presented simultaneously. This latter design option should involve the lowest working memory load and might therefore be considered particularly useful for the study of face-voice matching abilities in the absence of additional working memory-related load artifacts (as a consequence, this paradigm was utilized in Study 3 of this thesis: Huestegge, *subm.*).

Finally, there is another procedure that can be found in the literature. It involves a sequential display of two different pairs of face and voice. Then, participants should indicate which one of the two pairs was a match. This has been termed “same-different procedure” (Smith, Dunn, Baguley, & Stacey, 2016b), and corresponding results replicated the findings of successful matching abilities reported above. In a subsequent study, these authors demonstrated that the insertion of a delay between the display of the voice and the two faces eliminated matching abilities (Smith, Dunn, Baguley, & Stacey, 2018). This is first direct evidence for the idea that memory load indeed disrupts matching abilities. Another group replicated the effects by Smith et al. (2016b) by demonstrating above-chance matching using



static faces in a same-different procedure (Stevenage, Hamlin, & Ford, 2017). Furthermore, by comparing performance between the individual voice samples, they found that voice distinctiveness is associated with enhanced matching abilities.

#### **4. Theories of voice (and face) processing and implications of the present studies**

In this chapter, specific theories of voice (and face) processing will be introduced. It will become clear that theoretical reasoning about voice processing is historically closely linked to face processing theories, likely due to the fact that there is a large overlap of person-related types of information processed through both processing channels. Before specific theories will be described, some brief introductory remarks will highlight the vast aspects of information conveyed through voices.

##### **4.1. Introductory remarks: What's in a voice? General aspects of cognitive voice processing**

Many different speaker characteristics can principally be inferred from voices. For example, an unfamiliar person's height and age can be inferred from both voices and faces (Allport & Cantril, 1934; Lass & Colt, 1980). Of course, this also holds for a person's gender (e.g., Fellowes, Remez, & Rubin, 1997; Weston, Hunter, Sokhi, Wilkinson, & Woodruff, 2015). For example, a study that addressed gender classification based on auditory information was conducted by Meister, Kühn, Shehata-Dieler, Hagen, and Kleinsasser (2017). They had participants judge the gender of voices of transgender speakers via telephone-like acoustics and found that male-to-female individuals that were misperceived as male could be identified even when their vocal pitch was in the female frequency range.

Interestingly, recognizing gender from voices is also possible for children's voices prior to puberty despite a lack of known clear anatomical larynx and vocal tract differences at this early age (Fitch & Giedd, 1999; Sachs, Lieberman, & Erickson, 1973). Generally, men's voices have a lower average fundamental frequency and lower formant frequencies compared to women's voices, the latter typically exhibiting a higher low-to high frequencies ratio and increased aspiration noise (Hanson & Chuang, 1999; Klatt & Klatt, 1990; Titze, 1989). However, the ability to extract information from voices and faces also holds for personality

traits like extraversion and conscientiousness (Allport & Cantril, 1934; Berry, 1991; Borkenau & Liebler, 1992). Interestingly, even infants are able to match faces to voices based on both emotions (7 months, Walker-Andrews, 1986) and gender (8 months, Patterson & Werker, 2002). Taken together, a vast variety of person-related information can be extracted from voices alone.

#### **4.2 Voice (and face) processing theories**

Cognitive models of voice processing, including the specification of relevant cognitive structures, processing sequences and interconnections, owe much to previous progress in the field of face recognition. Therefore, the first models of voice processing (and of the interaction between face and voice processing) were primarily “copied” from the influential face information processing framework by Bruce and Young (1986). In this model, they capture the process of recognition of familiar faces by first assuming an initial structural face processing stage. This stage results in a viewpoint-invariant and expression-invariant face representation. Invariant representations are necessary to allow for object constancy, that is, to allow for perceiving the same face when this face is seen from various viewpoints. When faces are familiar, the model assumes that corresponding face recognition units are in place. These are activated by the presence of a corresponding invariant face representation to signal familiarity, a process assumed to be hampered in the case of prosopagnosia (see Section 3.1). At a subsequent stage, this information can additionally trigger retrieval of further semantic information about the person associated with the face. Only at a final stage, after the retrieval of semantic information, name retrieval is assumed to occur. The separation of a familiarity check and semantic information retrieval captures the experience of “having seen that face before” without being able to place (or finally to name) it. In parallel to this route of familiar face recognition, the model also assumes parallel routes dedicated to emotion processing, and one devoted to speech processing (relevant for lip reading). Finally, a fourth route was termed

“directed visual processing”, and proposed to be functionally linked to the processing of facial features such as age, gender, gaze etc. Despite the principal separation of the pathways, interaction is made possible by the connection of each pathway to the cognitive system (Bruce & Young, 2012).

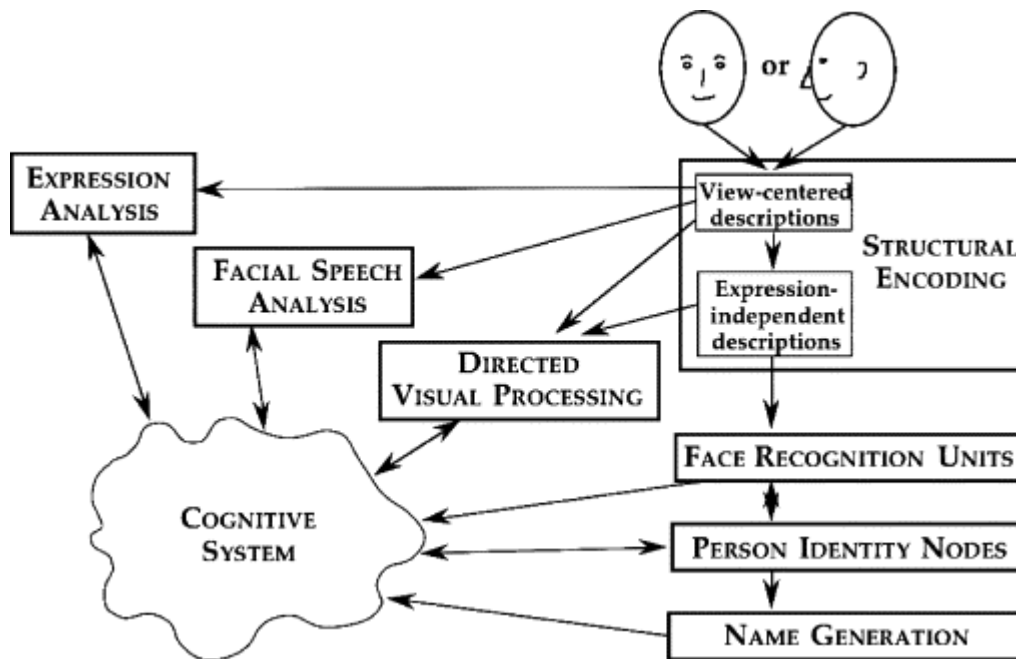


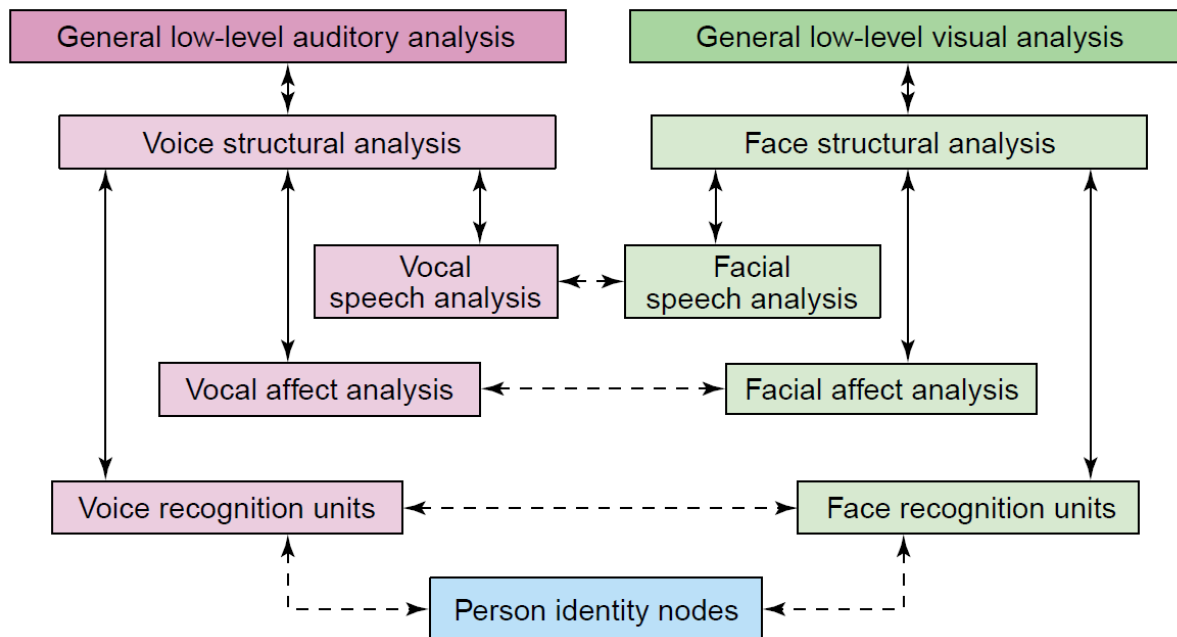
Figure 1. Face recognition model of Bruce and Young (1986), taken from: <https://sites.google.com/site/prosopagnosiafaceblindness/face-recognition>

After its initial introduction this model has been implemented computationally (Burton, Bruce, & Johnston, 1990), and it was then further extended to also incorporate voice recognition (Belin, Fecteau, & Bédard, 2004; Ellis, Jones, & Monsell, 1997; Stevenage, Hugill, & Lewis, 2012; Campanella & Belin, 2007), serving the overall goal of processing person-related information as a whole, for example, a person’s identity (in terms of recognizing a person).

However, these models have long been ambiguous regarding the question of the extent to which face and voice processing streams interact, that is, whether the two processing streams should be considered rather *modular* (i.e., able to operate largely independently as suggested by different underlying brain networks, see Stevenage & Neil, 2014; see Section 3.1), or whether the two streams of (face and voice) information processing should be regarded as essentially *integrative* due to the fact that they process very similar types of information (Campanella & Belin, 2007; see Section 3.2).

The most influential framework of voice processing (in the context of face processing) was proposed by Belin et al. (2004, see also Campanella & Belin, 2007). As a first stage, it postulates a low-level analysis of the auditory voice input, followed by a structural voice analysis (i.e., similar to the structural encoding stage in face processing) resulting in rather invariant, more abstract representations of the voice. On this basis three separate, parallel, but interactive processing pathways are assumed: A speech analysis pathway, an emotion processing pathway, and a voice recognition pathway, the latter involving voice recognition units which are activated in case of familiar voices (again, this is highly similar to corresponding assumptions in Bruce and Young's face processing model, see above). An active voice recognition unit is then able to trigger a corresponding person identity node that is shared with the face processing system.

All these pathways are assumed to produce increasingly abstract representations of the voice. The model not only assumes interaction between the aforementioned three voice processing pathways, but, importantly, also between corresponding pathways in face and voice processing. Note that the model has been developed based on neurophysiological plausibility, that is, each stage and pathway is supposed to be represented by a distinct set of neural correlates (based on ERP, fMRI, and MEG studies, among others) and brain activation pathways (see Belin et al., 2004, for details).



*Figure 2.* Interactive face-voice processing model involving three pathways, taken from Belin et al. (2004).

Surprisingly, Belin et al. (2004) as well as Campanella and Belin (2007) did not “copy” the fourth processing route referred to above when they transformed Bruce and Young’s (1986) model into a voice processing model, namely the route in which visual processing can be directed to specific features such as age, gaze, and gender (etc.). In the context of our present research questions (and results, see below), it appears wise to re-instantiate this route also for the case of voice processing.

#### **4.3 Situating the present research questions in the models**

In this chapter, the rationale of each of the three studies will be described in relation to the theoretical issues and modelling approaches that have been described so far.

**4.3.1 Study 1 (Huestegge & Raettig, in press): Gender-related face-voice processing interaction.** The Stroop-like experiment in Study 1 (Huestegge & Raettig, in press) regarding cross-modal gender-congruency effects on gender classification of faces and voices directly addresses the assumption of strong interaction between face and voice information on the level of structural face and voice analysis where gender information should be extracted. Specifically, we aimed to verify the postulated bi-directional interaction and to further test if this interaction is asymmetric, thereby potentially enriching the model. Additionally, we introduced a task switching design in order to test whether and how gender-congruency effects (if present) are modulated by the strength of executive control demands.

*Specific research questions and hypotheses:* In this study, we utilized a Stroop-like (in terms of the manipulation of the congruency between a relevant and an irrelevant stimulus dimension, see Stroop, 1935; MacLeod, 1991) dynamic interference task involving 288 gender-congruent or gender-incongruent synchronized audio-visual face videos uttering task-irrelevant digits. Participants indicated gender identity of either the face or the voice via a two-choice key press. Task (face vs. voice classification) was varied both in separate blocks of trials (associated with comparatively low executive control demands) and in a mixed block with unpredictably alternating tasks (associated with comparatively high executive control demands due to trial-by-trial switching of the two tasks). Color cues that were drawn around the face videos indicated the task to execute in each trial.

The four research questions regarding gender classification performance were as follows: 1. Does incongruent visual facial information negatively affect voice classification performance, and does incongruent voice information negatively affect visual face classification performance? The presence of such dynamic congruency effects would indicate strong bidirectional interaction between face and voice processing (Campanella & Belin, 2007), even for non-ambiguous gender information.

2. Is the negative effect of incongruent (vs. congruent) visual information on voice gender classification stronger than the corresponding effect of voice information on visual face gender classification? This would support theoretical claims postulating visual dominance in face-voice integration (see Stevenage & Neil, 2014, for a review; see Section 2.2.1).

3. Is the congruency effect (if present) mainly based on initial surprise associated with incongruent stimuli at the beginning of the experiment, or is it a temporally stable phenomenon? The latter would suggest a hard-wired, generic mechanism underlying audiovisual integration.

4. Is the congruency effect modulated by cognitive load? This question was addressed by including task-switching blocks associated with high executive control demands: In this condition, participants switch between visual face-based and voice-based gender classification from trial to trial in an unpredictable (random) manner (task switching paradigm, see Allport, Styles, & Hsieh, 1994, Rogers & Monsell, 1995). The task in each trial is indicated by the color of the frame drawn around the face video. Specifically, high executive processing demands are assumed to be associated with task switch trials, while task repetition trials should be associated with relatively lower executive demands (Kiesel et al., 2010). Thus, we were able to test whether congruency effects are modulated by trial type (task switches vs. repetitions). Two possible, complimentary outcomes were anticipated here: a) When higher executive control demands bind resources needed for resolving gender interference, one would expect *stronger* congruency effects in switch trials than in repetition trials. b) In contrast, it is also possible to assume that the increased binding of cognitive resources in switch (vs. repetition) trials leaves fewer resources dedicated to the processing of gender information in the processing channel which is irrelevant for the currently relevant gender classification task. In this case, relatively more attentional focus would be on the task-relevant



(vs. task-irrelevant) stimulus dimension), and we would therefore predict *weaker* congruency effects.

**4.3.2 Study 2 (Huestegge et al., in press): Impact of gender-congruency on basic, speech-based information processing.** Study 2 (Huestegge et al., in press), which involved exactly the same stimulus material as Study 1 (Huestegge & Raettig, in press), addressed effects of gender (in)congruency on digit processing, thereby focusing on effects of the interaction of combined face-voice gender processing on the pathway involving vocal speech analysis in models of voice processing. This pathway is required to extract the digit meaning, which represents a prerequisite for solving basic magnitude or parity tasks. Of course, digit processing here serves as a proxy for basic (speech-based) information processing in general.

*Specific research questions and hypotheses:* Overall, we propose that voices generally cannot be processed in isolation. Instead, they are always processed in conjunction with facial information (when available), eventually yielding easier processing for more authentic voices (i.e., voices which do not violate visual face-based expectancies). If this is true (and this is the central research question here), one would clearly expect to observe gender congruency effects on digit classification performance. As in Study 1 (Huestegge & Raettig, in press), we also implemented an additional cognitive load manipulation (in terms of task switching blocks) in Study 2 (Huestegge et al., in press). We therefore again explored the dependency of face-voice integrative processing on general cognitive load related to executive demands (tapping into central cognitive resources). This is especially interesting here, since unlike in Study 1 (Huestegge & Raettig, in press), the congruency operates on the level of task-irrelevant information (gender and gender congruency is not relevant for identifying/classifying the spoken digits). We also aimed at ensuring the stability of congruency effects over the course of the experiment to address the adaptability of integration processes in the model. Specifically, we asked whether we can get used to processing content

incongruencies, and whether we can adapt to previously unexperienced stimuli (male faces with female voices are usually not experienced often in daily life). If the corresponding effect is constant over the course of the experiment, this would indicate that gender congruency-based conflicts pose a persistent processing obstacle.

**4.3.3 Study 3 (Huestegge, *subm.*): Matching faces and voices.** Study 3 (Huestegge, *subm.*) represents a face-voice matching study that addressed the issue of common source information present in voices and faces as well as our cognitive access to them. Thus, successful matching behavior can be considered the outcome of learned contingencies between both processing pathways over lifelong experience with faces and voices. A central feature of this study is that a simultaneous presentation paradigm is used in order to avoid confounding effects of working memory load (see Section 3.2.4). Note that this study involves completely new stimulus material (compared to the previous studies).

*Background and specific research questions:* In this study, we designed an experiment involving both static and dynamic face conditions. Both conditions were integrated in a paradigm with fully simultaneous stimulus presentation, in which a voice is accompanied with the display of either two static faces (see Mavica & Barenholtz, 2013), or two dynamic faces, the latter representing a condition not examined until now in this paradigm, even though such a condition more realistically captures matching demands in daily life situations, such as when trying to find the face that called our name among many competing faces in a crowd.

Participants used left or right key press responses to indicate which of two faces presented to the left and right on a screen corresponds to a *simultaneously* presented voice. The stimulus models in the experiment were AV-synchronized so that they uttered a different text in the auditory stream than the text in the visual stream of the videos. Note that by presenting all information in both processing channels simultaneously we prevent any delay between a reference stimulus and comparison stimuli (see Section 3.2.4) that made

interpretations of previous studies somewhat difficult. Of course, a disadvantage of this procedure is that it is not feasible to include the modality-reversed condition, that is, a condition in which one visually presented face needs to be matched to either of two simultaneously presented voice streams.

The first main aim of Study 3 (Huestegge, *subm.*) was to replicate previous reports of the ability to match a voice to both static and dynamic faces, but by utilizing a new, simultaneous presentation paradigm which is similar to that utilized in a previous study on static face-voice matching (Mavica & Barenholtz, 2013). Thus, for the first time we test the ability to match both static *and* dynamic faces to a voice in a fully simultaneous presentation paradigm using the same underlying stimulus material (professional actors from the local theater in Würzburg). This enables us to clarify an open issue from previous research, namely the question of whether there really is an advantage of dynamic face-voice matching over static face-voice matching, as suggested, for example, by Kamachi et al. (2003). In order to reliably sort out any participants that did not pay sufficient attention to the matching task, we implemented an additional incidental learning paradigm that was aimed at determining inattentive participants. To sum up, the crucial research question here is whether dynamic face-voice matching is substantially better than static face-voice matching ability.

## **5. Results of the current studies and theoretical implications**

In this chapter, some of the main results of the three studies will be briefly described, and implications for face-voice processing theories will be highlighted, eventually culminating in an enhanced model of face-voice processing (originally introduced in Huestegge et al., in press) that takes into account the present main results from the three studies.

### **5.1 Results & implications of Study 1 (Huestegge & Raettig, in press)**

Taken together, the results in Study 1 (Huestegge & Raettig, in press) provide strong evidence, in both RTs and error rates, for gender congruency effects in terms of a bi-directional interaction of voice and face processing streams regarding gender information. This shows that even when participants try to focus their attention on one input channel only, this type of filtering based on a physical dimension (Broadbent, 1958) is not perfectly possible. Therefore, our data speak in favor of face-voice processing theories that assume a strong integration between face and voice processing. This integration is based on a common source of to-be-processed information, namely gender (see Campanella & Belin, 2007).

Previous models of voice processing (see above) only included three parallel processing pathways, namely those dealing with speech-based, emotion-based, and identity-based information (the latter serving the purpose of recognizing familiar persons). It is quite possible that this selection is not driven by the matter of research, but rather by the particular research interests of individual scientists or by prominent research topics in the field. The present results of Study 1 (Huestegge & Raettig, in press) suggest that such models may consequently be limited, and that they should be augmented with possibilities to include potential additional pathways, depending on the concrete task demands and processing needs of a person in a particular situation/environment. Here, the data suggest to include the possibility of a gender information processing pathway. In this way, the model is to some extent more similar to the original face-based processing model by Bruce and Young (1986),

since they included a “directed visual processing” possibility that captured several qualities of specific visual face processing needs.

Our data also support the idea of a bi-directional integration of visual and auditory processing routes even when the dynamic gender stimuli are unambiguous. Most previous studies were limited in that they a) studied gender congruency effects in a single direction of interference only, b) utilized gender-ambiguous stimuli, thereby limiting the scope of theoretical conclusions (Freeman & Ambady, 2011; Peynircioglu et al., 2017, Smith et al., 2007; see also Section 3.2.2), or c) used static instead of dynamic faces, the former being not suited to study natural, dynamic face-voice integration (Latinus et al., 2010; see Section 3.2.2). Thus, Study 1 (Huestegge & Raettig, in press) is the first study that allows us to conclude that voice processing and face processing *generally* interact bi-directionally. This aspect is captured in the newly developed model in Section 6 by including bi-directional arrows on the level of gender processing.

The generality of our findings is further supported by the observation that the gender congruency effect could also be observed in highly demanding situations involving substantial executive control demands (task switching; see Kiesel et al., 2010, for an overview), and by the observation that this effect continued to be present even after long experience with gender-incongruent stimuli over the course of the experiment, thereby ruling out a more simple explanation of the effect in terms of mere surprise associated with the first few incongruent stimuli displayed at the start of the experiment. From a more general viewpoint, these indications of the generality of the gender congruency effect speak against the strong cognitive modularity assumption in terms of encapsulated processing modules suggested by Fodor (1983). Furthermore, the data also speak against the claim that attention can be perfectly selective based on physical features such as input modality (see the concept of filtering by Broadbent, 1958).

Of course, given the many instances of cross-modal processing effects reviewed in the previous sections (in particular Section 3.2) it would have been implausible from the start (i.e., a theoretical strawman) to assume a *strong* modular account based on the ideas by Fodor (1983). However, it was still reasonable to assume a looser version of the modularity notion which only assumes that – given optimal conditions for filtering – participants are *principally able* to selectively activate only one processing route (i.e., only visual or only auditory). However, the present data suggest that this is very unlikely (at least as long as both sources of information are available). In sum, the ability to selectively activate only visual face-based or voice-based processing routes (as assumed by Ellis et al., 1997; Yovel & Belin, 2013) appears to be extremely limited.

Results from the task switching blocks also demonstrated that greater executive control demands in switch (vs. repetition) trials were associated with greater congruency effects. We explained this finding by assuming that fewer cognitive processing resources are available under high load, so that fewer resources are available to resolve cross-modal gender interference. A necessary precondition here could be that there is dimensional overlap regarding the task-relevant and task-irrelevant information, since both contained gender-related information. This assumption was directly tested (and corroborated) in Study 2 (Huestegge et al., in press; see Section 5.2)

Another interesting observation with implications for theory is that in pure blocks we found evidence of visual dominance: We observed a stronger effect of incongruent visual face information on voice categorization than vice versa. This is captured in the model in Figure 3 by using differential arrows (black vs. light grey) indicating the two directions of interaction between processing pathways. These observations are in line with previously reported instances of face processing dominance (over voice processing, see Latinus et al., 2010). One further relevant phenomenon here is the Facial Overshadowing effect (Cook & Wilding,

1997), which describes the observation that voice recognition is improved when voices are presented in isolation than in conjunction with their corresponding faces (see also Stevenage, Howland, & Tippelt, 2011; Stevenage, Neil, & Hamlin, 2014). Additionally, previous studies also indicated that voices are a comparatively weak cue to recognize the identity (or associated characteristics) of a person, at least when compared to using faces as a cue (e.g., Barsics & Bredart, 2012; Ellis, Jones, & Mosdell, 1997; Hanley, Smith, & Hadfield, 1998) and to semantics associated with a person.

## **5.2 Results & implications of Study 2 (Huestegge et al., in press)**

In Study 2 (Huestegge et al., in press), we also observed clear gender-congruency effects on digit classification performance. Note that unlike in Study 1 (Huestegge & Raettig, in press), congruency here was manipulated between two *task-irrelevant* stimulus dimensions, since participants did not have to process gender information in the first place to complete the task. Thus, the results make an even stronger point than Study 1 (Huestegge & Raettig, in press) for the assumption of mandatory integrative face-voice processing.

The data are not in line with an abstract view of information processing (see Section 2.1.2) which assumes that mainly task-relevant processing codes (and potentially also associated dimensional overlap of code sets) determine processing performance (Kornblum et al., 1990). The present findings of congruency effects between two task-irrelevant stimulus dimensions also go beyond observations from classic cognitive conflict tasks (such as Stroop or Flanker tasks, Stroop, 1935; Eriksen & Eriksen, 1974), since the latter always involve dimensional overlap between task-relevant and task-irrelevant stimulus dimensions (e.g., stimulus color, or letter category).

In contrast to abstract information processing accounts, the data from Study 2 (Huestegge et al., in press) are instead more in line with embodied views on cognition (e.g., Barsalou, 2008; Barsalou et al., 2003; see Wilson, 2002), which emphasize that cognitive

representations are strongly connected to their perceptual origins. For example, this aspect is captured in the Perceptual Symbol Systems account proposed by Barsalou (1999). This theory was explicitly presented as an alternative to a-modal theories postulating abstract cognitive representations (see Section 2.1). In particular, he assumed that cognitive representations are based on perceptual (instead of abstract) symbols, which are generated based on concrete perceptual experiences. During cognitive processes, it is assumed that perceptual symbols are implemented by means of association mechanisms that partially re-activate the sensory-motor origins in a top-down manner. From this perspective, one could argue that the mental representations of the digits in Study 2 (Huestegge et al., in press) are inextricably linked to the context of their perceptual origin. This contextual origin includes task-irrelevant facial and vocal features associated with the digits as well as their inter-relation (in terms of gender-based congruency). Of course, it is possible that the present effect necessitates that the two task-irrelevant stimulus dimensions are generally relevant to some extent for the participant in order to affect processing (e.g., gender information is essentially social information which is potentially relevant for any human being).

As mentioned above, our data from Study 2 (Huestegge et al., in press) again support face-voice processing accounts postulating strongly interactive (and relatively automatic) processing for different types of information including gender (e.g., Campanella & Belin, 2007; Yovel & Belin, 2013). This information processing can eventually also exert an influence on other pathways (here: speech processing, which is a precondition for solving the digit-related tasks). In our model Section 6, this is captured by a corresponding arrow from the gender processing pathway to the vocal speech analysis pathway. The fact that the present results speak in favor of relatively automatic processing (since gender information affected performance despite being task-irrelevant) appears to contradict (for the case of voice processing) a potential claim in Bruce and Young's face processing model, where a "directed visual processing" route was postulated. If we interpret this in terms of directed (controlled,



thereby non-automatic) attention, it seems as if our present results rather speak for the possibility of automatic processing of visual features, at least for certain types of features (e.g., socially relevant features).

One surprising observation in Study 2 (Huestegge et al., in press), and a result in contrast to Study 1 (Huestegge & Raettig, in press), was evidence for a *reversed* gender congruency effect in switch trials (as evident in RTs in the task switching block). As a reasonable potential post-hoc explanation for this finding, we reasoned that effects of processing prioritization or capacity scheduling might have played a role. The presence of a gender congruency effect itself (even when reversed) suggests the presence of sufficient capacity for computing gender-related congruency information. Probably, participants are aware that gender-incongruent switch trials are the most difficult trials in the whole experiment and therefore strategically assign especially high amounts of cognitive resources to these trials in particular. This eventually may have resulted in the paradoxical pattern of better performance in the most difficult trials (switch trials compared with the corresponding repetition trials). This potential explanation is further corroborated by a similar observation in another condition of the experiment: The magnitude task, which is easier than the parity task (see RTs and error rates in pure blocks), was associated with worse performance in a task switching context, an observation similar to one originally reported by Meuter and Allport (1999) in the context of language switching. Thus, it paradoxically appears easier to switch to the more difficult task than to switch to the easier task, a phenomenon that could also be explained by assuming that more resources are assigned to more difficult conditions (apart from other, inhibition-related explanations discussed by Meuter & Allport, 1999). Nevertheless, the reversed congruency effect should be replicated in a future study, and the particular explanation offered here should be tested more explicitly.

In sum, Study 2 (Huestegge et al., in press) indicated that congruency between task-irrelevant stimulus dimensions (voice-face gender congruency) modulated information processing regarding a third, task-relevant stimulus dimension (digit identity). This can also be taken as evidence for the claim that inauthentic voices (i.e., voices that are not expected based on displayed facial features) are more difficult to process than authentic (gender congruent) voices. For general theories of cognition, this result supports accounts assuming that mental representations are essentially grounded in their sensory origins and their situational contexts (situated cognition, Greeno, 1998).

### **5.3 Results & implications of Study 3 (Huestegge, subm.)**

Study 3 (Huestegge, subm.) introduced a novel procedure for face-voice matching that helps to determine whether participants payed attention to the task. This procedure is independent from merely judging matching performance (since chance-level matching performance can indicate both insufficient commitment of a participant *or* a lack of matching abilities). Specifically, this procedure was based on implicit learning support for matching (see manuscript for details). The most important result of this study was that there was no evidence for statistically superior matching performance with dynamic (vs. static) face stimuli, and there was no indication for any ceiling or floor effects. Given the small confidence interval for the difference, it can be concluded (despite the general difficulties associated with interpreting null effects) that at least no *meaningful* incremental usefulness of dynamic (vs. static) information in face-voice matching appears to exist. Notably, this conclusion contradicts earlier claims postulating that dynamic facial information is necessary (or at least highly relevant) for matching abilities (e.g., Kamachi et al., 2003).

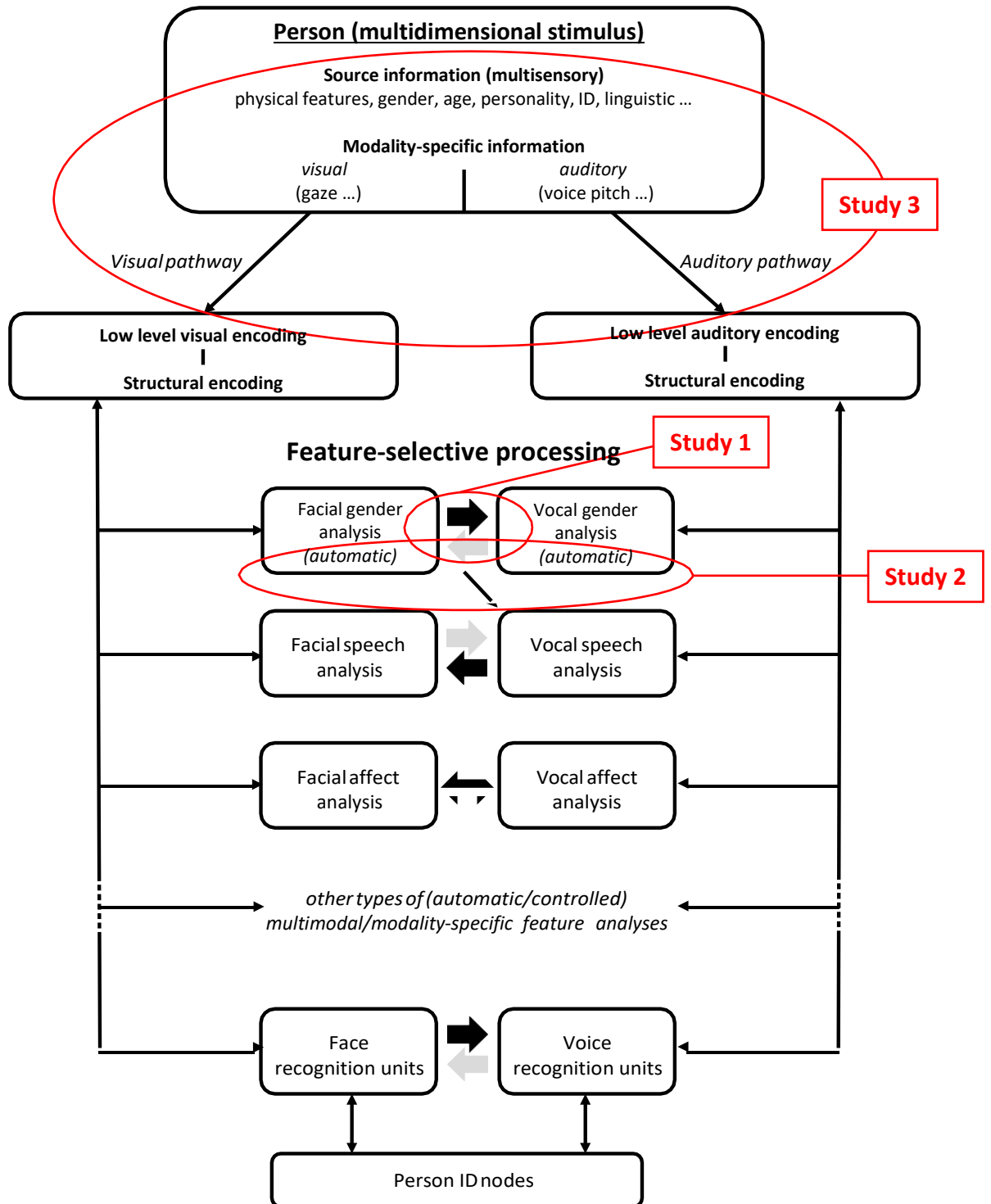
Within our model of face-voice processing (Section 6), Study 3 (Huestegge, subm.) addressed the fact that common person-related source information is present in both the face-based (visual) and the voice-based (auditory) processing stream. This is a precondition for any

ability of participants to infer a particular face from voice features (see upper circle depicted in Figure 3 in Section 6).

Note that Study 3 (Huestegge, *subm.*) is also the first study which directly compared static and dynamic faces in an ideal experimental situation for face-voice matching by simultaneously presenting *all* stimuli (auditory reference stimulus, visual comparison stimuli) to reduce any additional memory-related demands. As outlined in Section 4.3.3, sequential presentation paradigms (e.g., Lachs, 1999; Kamachi et al., 2003; Lachs & Pisoni, 2004 a,b; Lander et al., 2007) are not suited to fairly compare static and dynamic voice matching. Our observation of similar matching performance for static and dynamic faces suggests that whatever the source information used by participants to complete the task (e.g., hormone expression in both faces and voices etc., see Section 4.1), it should mainly be already present in the static face, and only to a negligible extent solely in the dynamic (speaking-related) visual information.

## **6. Final discussion: Novel theoretical insights**

Study 1 (Huestegge & Raettig, in press) addressed the issue of the interaction of gender-related face and voice information and demonstrated a processing asymmetry (visual dominance). Study 2 (Huestegge et al., in press) showed that this interactive processing is relatively automatic and can affect the processing of speech information. Finally, Study 3 (Huestegge, *subm.*) addressed the ability to match unfamiliar faces and voices and shows that common source information affects voices and faces, a phenomenon used by participants to be able to match faces to voices. Study 3 (Huestegge, *subm.*) further suggests that dynamic visual information is not crucial for matching, so that the relevant information must already be inherent in the static face. Overall, the results of all three studies support a strongly interactive, embodied (rather than a modular, abstract) view of cognitive processes in general (see Section 2). This conclusion is finally captured in a new model (Figure 3), which represents an account of integrative voice and face processing based on earlier suggestions (e.g., Belin et al., 2004), but which has been modified and augmented in order to be able to capture the results from the present set of studies (see Section 5).



*Figure 3.* Enhanced model of face-voice processing highlighting the contributions of the three studies reported in the present thesis (Huestegge, subm.; Huestegge & Raettig, in press; Huestegge et al., in press). This extended model was originally introduced in Huestegge et al. (in press).

Similar to previous models of either face processing (Bruce & Young, 1986) or face and voice processing (Belin et al., 2004), the input of the model is based on multi-modal stimulus information that inherently transports person-related information across visual and auditory modalities, such as age, personality, gender, etc. The fact that information is redundantly coded across these modalities (see Campanella & Belin, 2007) is essentially the reason why participants are able to match unfamiliar voices and faces (see Study 3: Huestegge, *subm.*), although it has remained an open question what kind of information exactly is the basis for these matching abilities. The findings of Huestegge (*subm.*) suggest that the relevant source information used for matching is already available in static faces.

In the model, modality-specific information is then processed downstream within two distinct, but interactive visual and auditory processing pathways. Similar to previous face and/or voice processing models (Belin et al., 2004; Bruce & Young, 1986), it is assumed that in each of these pathways, an initial low-level encoding eventually results in structural representations, which are then subject to more specific, feature-selective processing devoted to different types of information (e.g., gender). The model also adopts the idea of including face-based and voice-based recognition units (to recognize a familiar person, see bottom of the model in Figure 3) from previous models (e.g., Belin et al., 2004), even though this particular issue is not addressed in the present set of studies.

This new model of integrative voice and face processing (Figure 3) is characterized by four important new features. First, it generally allows for more than three processing pathways (see Belin et al., 2004, for a model with only three routes). In particular, it also includes a dedicated gender processing pathway to be able to account for the gender congruency effects found in Study 1 (Huestegge & Raettig, *in press*) and Study 2 (Huestegge et al., *in press*). Generally, a key feature of the present model is that it is explicitly open to

account for any conceivable additional type of information (in terms of feature-selective processing) that can be present in voices and faces (captured by the dotted lines towards other types of multimodal person-related features). Thus, depending on specific situational demands or individual processing needs/expertise, corresponding pathways could be assembled on the fly.

Second, the model is also able to account for cross-modal processing asymmetries within different pathways. This asymmetry is captured in the model by the differential arrow colors for both directions of influence (i.e., from visual to auditory processing and vice versa). Note that these asymmetries are not assumed to be constant for all types of information: In line with our results from Study 1 (Huestegge & Raettig, in press), visual dominance in the case of gender processing was postulated. This also holds for processing a person's identity in terms of recognizing a familiar person: Previous studies have shown that recognition abilities are better for (visual) faces than for (auditory) voices (e.g., Barsics & Bredart, 2012; Ellis et al., 1997; Hanley et al., 1998). However, this is certainly different in speech processing: To understand the content of speech, the dominant modality is the auditory stream of information, which, however, can be modified by visual lip reading (as demonstrated by McGurk & MacDonald, 1976). Thus, in this case the black arrow (depicting greater influence than the light grey arrow) points from the auditory towards the visual stream. The model does not incorporate an asymmetry for affect analysis, since in this case more research is needed to decide whether affect processing is generally more dominant in the visual or in the auditory domain (e.g., see Schirmer & Adolphs, 2017).

Third, our model additionally allows for specifying each pathway in terms of its level of automaticity. As outlined in Section 5.2, the original face processing model by Bruce and Young (1986) involved a "directed visual processing" route, suggesting a role of directed (controlled, thereby non-automatic) attention for processing specific features of a face (e.g.,

gender). However, Study 2 (Huestegge et al., in press) instead indicates, at least for the case of combined face and voice processing, the possibility of automatic processing of the (task-irrelevant) gender relation between visual and auditory processing streams. Of course, future research is needed to determine the level of potential processing automaticity for the various other types of information (speech, facial affect etc.) in the model.

Fourth, the model also allows for particular interactions between bi-modal processing of different types of information. This was not explicitly captured by previous models. Specifically, based on the results of Study 2 (Huestegge et al., in press), an influence of (task-irrelevant) bi-modal gender processing on speech analysis was identified, the latter representing a necessary precondition to solve the digit categorization tasks. Future research should be devoted to systematically assess other types of interactions between processing routes for different types of information.

In sum, the present modified and extended model of voice and face processing is able to account for a majority of effects found in the present set of studies (Huestegge, *subm.*, Huestegge & Raettig, *in press*; Huestegge et al., *in press*) as well as in research from other groups. In addition, it also has the potential to stimulate various new lines of research, which should be subject of future research.



## 7. References

- Allport, G. W., & Cantril, H. (1934). Judging personality from voice. *The Journal of Social Psychology, 5*, 37-55.
- Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance series. Attention and performance 15: Conscious and nonconscious information processing* (pp. 421-452). Cambridge, MA: MIT Press.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-609.
- Barsalou L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-45.
- Barsalou, L. W., Simmons, W. K., Barbey, A., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*, 84-91.
- Barsics, C., & Brédart, S. (2012). Recalling semantic information about newly learned faces and voices. *Memory, 20*, 527-534.
- Barton, J. J. (2008). Structure and function in acquired prosopagnosia: lessons from a series of 10 patients with brain damage. *Journal of Neuropsychology, 2*, 197-225.
- Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*, 129-135.
- Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*, 535-543.
- Berry, D. S. (1991). Accuracy in social perception: Contributions of facial and vocal information. *Journal of Personality and Social Psychology, 61*, 298-307.
- Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review, 5*, 482-489.
- Boltz, M. G. (2017). Facial biases on vocal perception and memory. *Acta Psychologica, 177*, 54-68.

- Borkenau, P., & Liebler, A. (1992). The cross-modal consistency of personality: Inferring strangers' traits from visual or acoustic information. *Journal of Research in Personality, 26*, 183-204.
- Broadbent, D. E. (1958). *Perception and communication*. New York: Oxford University Press.
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305-327.
- Bruce, V., & Young, A. (2012). *Face perception*. New York: Psychology Press.
- Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97*, 523-547.
- Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81*, 361-380.
- Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*, 535-543.
- Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition, 26*, 780-790.
- Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics, 16*, 409-412.
- Cook, S., & Wilding, J. (1997). Earwitness Testimony 2: Voices, Faces and Context. *Applied Cognitive Psychology, 11*, 527-541.
- de Gelder, B. & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion, 14*, 289-311.
- Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88*, 143-156.
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon identification of a target letter in a non- search task. *Perception and Psychophysics, 16*, 143-149.
- Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics, 59*, 839-849.

- Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, *106*, 1511-1522.
- Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.
- Fodor, J. A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.
- Freeman, J. B., & Ambady, N. (2011). When two become one: Temporally dynamic integration of the face and voice. *Journal of Experimental Social Psychology*, *47*, 259-263.
- Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clinical Neurophysiology*, *120*, 453-463.
- Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, *50*, 524-536.
- Greeno, J. G., & Middle School Mathematics through Applications Project Group (1998). The situativity of knowing, learning, and research. *American Psychologist*, *53*, 5-26.
- Hagan, C. C., Woods, W., Johnson, S., Calder, A. J., Green, G. G., & 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 of the United States of America*, *106*, 20010-20015.
- Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *Quarterly Journal of Experimental Psychology*, *51*, 179-195.
- Hailstone, J. C., Crutch, S. J., Vestergaard, M. D., Patterson, R. D., & Warrena, J. D. (2010). Progressive associative phonagnosia: A neuropsychological analysis. *Neuropsychologia*, *48*, 1104-1114.

- Hanson, H. M., & Chuang, E. S. (1999). Glottal characteristics of male speakers: acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America*, *106*, 1064-1077.
- Herald, S. B., Xu, X., Biederman, I., Amir, O., & Shilowich B. E. (2014). Phonagnosia: A voice homologue to prosopagnosia. *Visual Cognition*, *22*, 1031-1033
- Hietanen, J. K., Leppänen, J. M., Illi, M., & Surakka, V. (2004). Evidence for the integration of audiovisual emotional information at the perceptual level of processing. *European Journal of Cognitive Psychology*, *16*, 769-790.
- Huestegge, S., & Raettig, T. (in press). Crossing gender borders: Bidirectional dynamic interaction between face-based and voice-based gender categorization. *Journal of Voice*. <https://doi.org/10.1016/j.jvoice.2018.09.020>
- Huestegge, S. M., Raettig, T., & Huestegge, L. (in press). Are face-incongruent voices harder to process? Effects of face-voice gender incongruency on basic cognitive information processing. *Experimental Psychology*.
- Joassin, F., Maurage, P., & Campanella, S. (2011). The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study. *NeuroImage*, *54*, 1654-1661.
- Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology*, *13*, 1709-1714.
- Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, *136*, 849-874.
- Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, *87*, 820-857.

- Knappmeyer, B., Thornton, I. M., & Bühlhoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, *43*, 1921-1936.
- Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility – A model and taxonomy. *Psychological Review*, *97*, 253-270.
- Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, *38*, 618-625.
- Lachs, L. (1999). A voice is a face is a voice: Cross-modal source identification of indexical information in speech. In *Research on spoken language processing* (Progress Report No. 23, pp. 241-258). Bloomington, IN: Indiana University, Department of Psychology, Speech Research Laboratory.
- Lachs, L., & Pisoni, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology*, *16*, 159-187.
- Lachs, L., & Pisoni, D. B. (2004b). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*, *116*, 507-518.
- Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*, 429-442.
- Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: Matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 905-914.
- Lass, N. J., & Colt, E. G. (1980). A comparative study of the effect of visual and auditory cues on speaker height and weight identification. *Journal of Phonetics*, *8*, 277-285.
- Latinus, M., VanRullen, R., & Taylor, M. J. (2010). Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC Neuroscience* *11* (11), 36.

- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin, 109*, 163-203.
- Marx, M. H., & Cronan-Hillix, W. A. (1987). *Systems and theories in psychology (4th ed.)*. New York: McGraw-Hill Book Company.
- Masuda, S., Tsujii, T., & Watanabe S. (2005). An interference effect of voice presentation on face gender discrimination task: Evidence from event-related potentials. *International Congress Series, 1278*, 156-159.
- Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 307-312.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*, 375-407.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.
- Meister, J., Kühn, H., Shehata-Dieler, W., Hagen, R., & Kleinsasser, N. (2017). Perceptual analysis of the male-to-female transgender voice after glottoplasty – The telephone test. *Laryngoscope, 127*, 875-881.
- Meuter, R. F. I., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language, 40*, 25-40.
- O'Toole, A. J., Roark, D., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science, 6*, 261-266.
- Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81*, 93-115.

- Peynircioglu, Z., Brent, W., Tatz, J., & Wyatt, J. (2017). McGurk effect in gender identification: Vision trumps audition in voice judgments. *The Journal of General Psychology, 144*, 59-68.
- Pourtois, G., de Gelder, B., Bol, A., & Crommelinck, M. (2005). Perception of facial expressions and voices and of their combination in the human brain. *Cortex, 41*, 49-59.
- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207-231.
- Sachs, J., Lieberman, P., & Erickson, D. (1973). Anatomical and cultural determinants of male and female speech. In: R. Shuy, & R. W. Fasold (eds.), *Language attitudes: current trends and prospects* (pp. 74–84). Washington, DC: Georgetown University Press.
- Schirmer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch: Comparisons and convergence. *Trends in Cognitive Sciences, 21*, 216-228.
- Schweinberger, S. R., Kloth, N., & Robertson, D. M. C. (2011). Hearing facial identities: Brain correlates of face–voice integration in person identification. *Cortex, 47*, 1026-1037.
- Shams, L., Kamitani, Y. & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14*, 147-152.
- Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016a). Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics, 78*, 868-879.
- Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016b). Concordant cues in faces and voices: Testing the back-up signal hypothesis. *Evolutionary Psychology, 14*, 1474704916630317.

- Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2018). The effect of inserting an inter-stimulus interval in face–voice matching tasks. *The Quarterly Journal of Experimental Psychology*, *71*, 424-434.
- Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology*, *17*, 1680-1685.
- Spence, C. (2009). Explaining the Colavita visual dominance effect. *Progress in Brain Research*, *176*, 245-258.
- Stevenage, S., Hamlin, I., & Ford, B. (2017). Distinctiveness helps when matching static faces and voices. *Journal of Cognitive Psychology*, *29*, 289-304.
- Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology*, *25*, 112-118.
- Stevenage, S. V., Hugill, A., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *Journal of Cognitive Psychology*, *24*, 409-419.
- Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica*, *54*, 266-281.
- Stevenage, S. V., Neil, G. J., & Hamlin I. (2014). When the face fits: recognition of celebrities from matching and mismatching faces and voices. *Memory*, *22*, 284-94.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.
- Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, *85*, 1699-1707.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.
- Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: a dissociation between familiar and unfamiliar voices. *Cortex*, *24*, 195-209.



- Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relation of eye and voice? *Developmental Psychology, 22*, 373-377.
- Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory “compellingness” in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics, 30*, 557-64.
- Weston, P. S., Hunter, M. D., Sokhi, D. S., Wilkinson, I. D., & Woodruff, P. W. (2015). Discrimination of voice gender in the human auditory cortex. *Neuroimage, 15*, 105, 208-14.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review 9*, 625-636.
- Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1*, 202-238.
- Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Science, 17*, 263-271.
- Zweig, L. J., Suzuki, S., & Grabowecky, M. (2015). Learned face-voice pairings facilitate visual search. *Psychonomic Bulletin & Review, 22*, 429-436.

# Study 1

Huestegge, S., & Raettig, T. Crossing gender borders: Bidirectional dynamic interaction between face-based and voice-based gender categorization

In press: *Journal of Voice* (accepted for publication on 25.09.2018)  
Published "online first": <https://doi.org/10.1016/j.jvoice.2018.09.020>

Bei Studie 1 handelt es sich um eine Version (post-review, pre-proof) des o.g. Artikels (Nutzung hier mit freundlicher Genehmigung des Elsevier-Verlags).

# Study 2

Huestegge, S. M., Raettig, T., & Huestegge, L. (2019). Are face-incongruent voices harder to process? Effects of face-voice gender incongruency on basic cognitive information processing. *Experimental Psychology*, 66, 154-164. (accepted for publication on 10.01.2019)

Bei Studie 2 handelt es sich um eine Version (post-review, pre-proof) des o.g. Artikels (Nutzung hier mit freundlicher Genehmigung von *Experimental Psychology* 2019; Vol. 66(2), 154–164 ©2019 Hogrefe Publishing; [www.hogrefe.com](http://www.hogrefe.com)).

DOI: <https://doi.org/10.1027/1618-3169/a000440>

# Study 3

Sujata M. Huestegge (subm.). Matching unfamiliar voices to static and dynamic faces: No evidence for a dynamic face advantage in a simultaneous presentation paradigm

Submitted: *Frontiers: Psychology* (15.02.2019)

Bei Studie 3 handelt es sich hier um eine Version (initially submitted version, but without any changes that occurred during the revision process) des o. g. Artikels.

Die finale Version wurde am 08.08.2019 von *Frontiers: Psychology* akzeptiert und kann unter DOI: 10.3389/fpsyg.2019.01957 (Copyright © 2019 Huestegge) abgerufen werden.

**Crossing gender borders:  
Bidirectional dynamic interaction between face-based and voice-based gender  
categorization**

Sujata M. Huestegge<sup>1,2</sup> & Tim Raettig<sup>1</sup>

<sup>1</sup>University of Würzburg

<sup>2</sup>University of Music and Performing Arts Munich

Correspondence address:

Sujata M. Huestegge

University of Music and Performing Arts Munich

Arcisstr. 12

80333 Munich, Germany

[sujata.huestegge@hmtm.de](mailto:sujata.huestegge@hmtm.de)

## Abstract

The processing of voices and faces is known to interact, for example, when recognizing other persons. However, few studies focus on both directions of this interaction, including the influence of incongruent visual stimulation on voice perception. In the present study, we implemented an interference paradigm involving 1152 videos of faces with either gender-congruent or gender-incongruent voices. Participants were asked to categorize the gender of either the face or the voice via key press. Task (face-based vs. voice-based gender categorization task) was manipulated both block-wise (relatively low executive control demands) and in a mixed block (relatively high executive control demands due to trial-by-trial task switches). We aimed at testing whether and how gender-incongruent stimuli negatively affected gender categorization speed and accuracy. The results indicate significant congruency effects in both directions – gender-incongruent visual information negatively affected voice categorization time and errors, and gender-incongruent voices affected visual face categorization. However, the former effect was stronger, supporting theories postulating visual dominance in face-voice integration. Congruency effects, which were not significantly reduced over the course of the experiment, were larger under high executive control demands (task switches), suggesting the availability of fewer attentional resources for incongruency resolution. Overall, voices generally appear to be processed in conjunction with facial information, which yields enhanced processing for more authentic voices, that is, voices that do not violate face-based expectancies. The data strengthen theories of face-voice processing emphasizing strong interaction between both processing channels.

Keywords: Voice Gender Processing; Stroop-like effect; Crosstalk; Face-voice integration; Gender; Visual dominance; Cognitive load; Auditory-visual interference

## 1. Introduction

*1.1 Processing voices and faces: Modular vs. interactive accounts.* When orally communicating with others in daily life, we usually do not process auditory information in isolation, but rather integrate streams of information from multiple channels simultaneously, including information from the visual processing channel. Two views can be distinguished regarding the underlying cognitive architecture of audio-visual person perception: According to a strongly *modular* account of cognition [1], face and voice information should be processed largely independently. For example, already Broadbent [2] suggested that people should be able to easily filter their attention based on (e.g., visual and auditory) modality. Evidence for such a modular organization comes from functional neuroanatomy indicating separate brain networks involved in the processing of faces (e.g., fusiform face area) and voices (e.g., temporal voice area) [3], and from neuropsychological patients: Some patients show a severe impairment of visual face processing (prosopagnosia), but their ability to process other objects or to recognize voices is still intact [4]. Conversely, patients suffering from phonagnosia are not able to recognize individuals based on their voice while face and name processing can remain unimpaired [5, 6, 7, 8]. Accordingly, a dominant theoretical framework of face processing [9], which has been augmented to incorporate voice recognition [10], in essence postulated parallel but rather separate face/voice processing pathways, that is, pathways that can principally be activated selectively [3]. Therefore, these pathways were originally assumed to lack strong interaction [11] within a general model of person perception [12].

However, there is also evidence for a strongly *interactive* account of visual and auditory processing, as evidenced by integration and crosstalk phenomena between both processing channels. This has led to theoretical frameworks of face-voice processing that strongly emphasize integrative processes based on the similarities of to-be-processed

information in both channels [13]. For example, research on cross-modal attention and multisensory integration has shown that when auditory stimulation is accompanied by a visual stimulus, auditory processing can be strongly attenuated up to the point of complete neglect (Colavita visual dominance effect [14, 15]). In a similar vein, the ventriloquism effect implies that sound localization is biased towards the location of simultaneously presented visual objects [16, 17, 18]. Other researchers have focused on the probably most socially important types of stimuli in the auditory and visual domains, namely voices and faces. Both convey information related to a person's age, affective state, personality, speech content, familiarity, and other person-related characteristics. Notably, both voices and faces are usually processed at the same time, thereby enabling integration [19, 20] or interaction in terms of facilitatory processing in the case of congruent information and interference in the case of incongruent information.

The probably most prominent audio-visual integration effect is related to speech information: McGurk and MacDonald [21] showed audiovisual faces combining, for example, the auditory syllables /ba-ba/ with visual synchronized lip movements of /ga-ga/. In this case, participants reported to hear /da-da/ (i.e., a fused illusory percept). Another study [22] reported that the McGurk effect can still be observed when the gender of the voice and the visual face did not match, implying that both streams of information came from different speakers. Therefore, audiovisual integration does not necessitate the perception of an integrated source (a single person) of both inputs. Apart from speech information, audio-visual integration in terms of congruency effects has also been shown for affective state (emotion) information [23, 24, 25, 26], age [27], and familiarity of a person [28]. Additionally, it was shown that the presence of a corresponding voice facilitates visual search for a particular face [29]. Together, these studies demonstrate that audiovisual integration in person perception occurs with respect to various types of information. In the following, we will focus on the more specific issue of processing gender-related information<sup>1</sup>.

*1.2 Integrative processing of gender information?* Audio-visual interaction effects were also reported for gender information. For example, one study [30] showed that audiovisual face presentation (i.e., when the voice of a person was presented along with the person's face) speeded up gender categorization when compared with a purely visual condition, whereas auditory presentation in isolation yielded slowest gender categorization performance. Overall, gender categorization has mostly been studied on the basis of *either* visual *or* vocal stimulus information, with a majority of studies focusing on the former (i.e., only visual face gender has to be categorized). For example, Smith, Grabowecky, and Suzuki [31] aimed to test how sounds that refer to gender speech are integrated with facial cues. They showed that gender-ambiguous (androgynous) faces were more likely to be categorized as male when presented with pure tones of low (vs. high) frequency (and vice versa for female faces). The authors assumed that the low/high tones represent the male/female fundamental speaking frequency range. Another study [32] demonstrated that gender-incongruent (as opposed to congruent) voices can disrupt the processing of facial gender, as evident in a modulated N170, an event-related potential closely linked to face processing. Freeman and Ambady [33] utilized (originally male and female) faces that were morphed so that their gender was ambiguous. In addition, they presented either original male/female (sex-typical) voices or voices with altered formants so that they became sex-atypical. Their results also showed that the voice information biased the participants' response to the faces: A sex-atypical voice attenuated the face gender categorization process.

Until now, only little is known about the impact of visual face information on voice gender discrimination (one previous study focused on voice gender discrimination in isolation [34]), maybe because there are generally fewer studies of voices than of faces [3]. However, a recent study [35] focused on a related issue, namely the effects of perceived visual gender on voice type categorization (bass, baritone, alto, soprano) using stimuli consisting of notes sung by male and female singers at the same g3 pitch. Results of categorization solely based on the

auditory tracks revealed that one third of the participants were not able to correctly categorize these auditory stimuli, suggesting that they were somewhat ambiguous with respect to voice type identity, likely due to the fact that pitch – an important voice type *and* voice gender cue – was held constant. In the crucial comparison, these auditory stimuli were visually synchronized with either a male or a female singer. Subsequent voice type categorization performance (based on only two audio-visual 1s-clips) was biased towards the visual gender information that was presented alongside (e.g., a male voice singing g3 was more likely to be categorized as alto/soprano when presented with a female face).

From our perspective, it appears likely that especially for the somewhat ambiguous stimuli used in the aforementioned studies participants may turn to any accessory information (in the other channel) that could help to disambiguate the percept in the light of the response options at hand. Thus, previous studies on audio-visual integration of gender information are still compatible with the general notion of modular, independent processing under *normal* conditions, that is, when both face and voice gender information is not ambiguous.

To sum up, none of the previous studies on gender categorization during face-voice integration outlined above involved a design suited to assess the bidirectional influence between voice and visual face gender information, and none of these studies involved non-ambiguous gender information in both processing channels, the latter establishing a situation in which the emergence of congruency effects cannot be simply ascribed to the reliance on external channel cues for the special case of disambiguation purposes.

There is only one previous study that addressed this research gap by focusing on bidirectional congruency effects on face and voice gender categorization [36]. In this study, gender-congruent and gender-incongruent static photos were presented alongside male/female voices uttering monosyllabic (French) words in a block-wise design (separate blocks for face vs. voice categorization). Interference effects were found in both directions, but there was



stronger (visual) face on (auditory) voice interference than vice versa. However, by using static photos this modality asymmetry of the congruency effect is confounded with a modality asymmetry in temporal presentation conditions in this study: Since the static faces did not move alongside the (inherently dynamic) voice presentation, this might have drawn special attention to the (non-moving) faces, potentially contributing to the observed visual dominance. Thus, a relevant precondition for studying face-voice *integration*, namely the synchronized dynamic presentation of speech in both modalities (as demonstrated by McGurk & MacDonald [21]), is not met. One aim of our present study is to close this gap by studying bidirectional gender congruency using natural dynamic (synchronized) videos in order to more directly address face-voice integration processes.

*1.3 The present study.* In the present study, we implemented a (Stroop-like [37, 38]) dynamic face-voice gender interference paradigm to determine the presence and extent of bidirectional integrative interaction between face-based and voice-based gender processing. Participants were presented with short videos of faces with either gender-congruent or gender-incongruent voices (uttering task-irrelevant digits) and were asked to categorize the gender of either the face or the voice via key press. Task (i.e., face-based vs. voice-based gender categorization) was manipulated both block-wise (relatively low executive control demands) and in a mixed block (relatively high executive control demands due to trial-by-trial task switching), the latter involving color cues indicating the task.

We focused on four hypotheses: 1. Based on the literature on audio-visual integration effects reviewed above, we predicted that gender-incongruent visual information should negatively affect voice categorization performance, and gender-incongruent voice information should negatively affect (visual) face categorization performance (Hypothesis 1). The presence of such bidirectional dynamic congruency effects, which have not yet been addressed in previous research, would indicate strong bidirectional integration between face

and voice processing streams [13] even for non-ambiguous gender information. 2. Based on previous theories postulating visual dominance in face-voice integration [12] we predicted that the negative effect of incongruent (vs. congruent) visual information on voice gender categorization should be stronger than the corresponding effect of voice information on visual face gender categorization (Hypothesis 2). 3. Previous literature on statistical learning suggests that behavioral responses to combinations of stimulus features which were previously never (or infrequently) experienced are severely impaired [39]. Since the gender-incongruent stimuli used in the present experiment also represent such novel (potentially surprising) feature combinations, especially at the beginning of the experiment, it is possible that any congruency effect is only driven by particularly slow (or error-prone) responses in incongruent trials at the beginning of the experiment, whereas increasing familiarity with gender-incongruent stimuli over the course of the experiment should lead to a decrease of the congruency effect with increasing time-on-task (Hypothesis 3). This possibility has not yet been considered in previous gender congruency studies, although it represents an important potential alternative explanation of the congruency effect. 4. Finally, by including blocks with high executive demands in which participants have to switch between visual face-based and voice-based gender categorization on a trial-by-trial basis (task switching [40,41]), we test (for the first time) whether high executive processing demands (associated with task switch trials as opposed to task repetition trials in such mixed blocks [42]) modulate the congruency effects (see a related study of load effects on audiovisual speech recognition [43]). On the basis of previous research [44], two potential outcomes appear conceivable: First, one could expect *stronger* congruency effects in switch (vs. repetition) conditions (Hypothesis 4a), because higher cognitive demands necessary to complete a switch trial [42] might consume resources needed for resolving interference based on gender incongruency. Second, it is also possible to expect *attenuated* congruency effects in switch (vs. repetition) trials (Hypothesis 4b), since the increased consumption of cognitive resources for the completion of switch trials

may yield a decrease of resources dedicated to the processing of task-irrelevant information (i.e., less attentional focus on information in the channel which is not relevant for the categorization task [45]).

## 2. Method

*2.1 Participants.* Twenty-four participants took part in the experiment (5 male, mean age = 23 years,  $SD = 3.8$ , range = 18-32). All participants had normal (self-reported) hearing and normal or corrected-to-normal vision. They received monetary reimbursement (or a small present) for participation. All participants gave informed consent.

*2.2 Stimuli & Apparatus.* Participants were seated about 67 cm in front of a TFT computer screen (15'') with a standard computer keyboard in front of them. Two keyboard keys (left Ctrl, Alt) served as response keys to indicate gender of faces or voices. Stimuli were short (fixed duration of 1 s; 25 frames) video clips of male or female faces uttering German digits (1, 2, 3, 4, 6, 7, 8 or 9) with a synchronized male or female voice (presented via headphones). The stimuli were based on original video recordings of three male and three female speakers (professional actors) uttering the digits. Since we aimed at natural visual face presentation with clear gender cues, the videos also showed the hair and upper torso of the actors. The six actors (dialect-free native speakers, neutral facial expression) were selected to ensure that both their visual appearance as well as their voice was unambiguously male or female. This was validated by informally presenting either the faces or the voices (uttering the digits) to 12 participants (no overlap with the sample from the main experiment), who correctly (100%) categorized all six faces and voices in terms of their gender. Male and female actors were (within reasonable limits) matched regarding their age. Specifically, one male/female pair was aged 35/32 years, one pair 41/37 years, and one pair 53/48. To rule out any potential advantage associated with unsynchronized (original) videos (which could only occur in the congruent condition), all videos were synchronized (using Pinnacle 21 software),

that is, for all stimuli the heard voice was never that of the original video. However, the stimuli also included self-synchronized videos, that is, videos in which the voice sample from one take of a participant uttering a digit was synchronized with the visual sample of another take of the same participant. It was ensured that voice information (equivalent to the visual face information) was immediately present in the first frame (although voice information may take some time to allow for gender information extraction). A visual fade out was implemented at the end of each video (final 8 frames) to ensure a smooth visual experience. In half of the stimuli the gender of the voice matched with that of the face, whereas the other half consisted of incongruent face-voice pairings. Altogether, these voice-face pairings resulted in  $8 \text{ (digits)} * 36 \text{ (voice-face pairings)} = 288$  stimuli (half of them congruent vs. incongruent regarding face/voice gender) that were presented in the middle of the screen ( $22^\circ * 12^\circ$  visual angle) on black background.

*2.3 Procedure.* The experiment (using the software Presentation®, Version 19.0, Neurobehavioral Systems, Inc., Berkeley, CA, [www.neurobs.com](http://www.neurobs.com)) started with an instruction screen. There were four blocks, two pure blocks and two mixed blocks. Participants in the pure blocks (i.e., blocks in which either only the voices or only the faces of the dynamic audiovisual stimuli should be categorized with respect to gender, while the other stimulus dimension had to be ignored) were asked to attend to the video clips (preceded by a central fixation cross presented for 250 ms) and to respond as fast and accurately as possible by pressing the left key for “female” and the right key for “male”. They were additionally instructed to look at the videos even when engaged in voice categorization. Task order of the pure blocks (visual face gender categorization task, auditory voice gender categorization task) was counterbalanced across participants. Each of the 288 stimuli was shown twice during each half of the pure blocks (i.e., for each of the two tasks). Each video stimulus was surrounded by a red or green rectangle (in randomized sequence). These were only presented for the sake of comparison with the mixed block condition (see below) and participants were

instructed to ignore them in the pure blocks. The two mixed blocks (also involving 2\*288 stimuli) consisted of random trial-by-trial switches between the face-based and the voice-based gender categorization task (e.g., a switch trial involving face gender categorization was preceded by a voice gender categorization trial, whereas a repetition trial involving face gender categorization was preceded by a face gender categorization trial). The task (in terms of the dimension which should be categorized) in each trial was indicated by a red vs. green rectangle (cued task switching [46]) surrounding the face video (color-task assignment was made explicit in the instructions and was counterbalanced across participants). Block order (pure blocks or mixed blocks first) was counterbalanced across participants. All stimuli were presented in random order within each block. The inter-trial interval was constant (100 ms). The experiment lasted about 40 minutes.

*2.4 Design.* Reaction times (RTs) and error rates (%) served as dependent variables. Separate ANOVAs for each dependent variable and for each block type (pure vs. mixed) were computed. The pure blocks analysis involved the within-subject factors congruency (congruent vs. incongruent face-voice pairings) and task (referring to the dimension of the stimuli which should be categorized: face vs. voice gender categorization task). The mixed block analysis additionally involved the factor sequence type (switch vs. repetition, i.e., whether the trial represented a task switch vs. repetition when compared with the previous trial). To test whether congruency effects are modulated by time-on-task, we additionally divided trials into four consecutive quarters (in pure blocks and in mixed blocks) and tested for significant interactions of congruency with quarter (as an additional factor).

### **3. Results**

*3.1 Data treatment.* For the RT analyses, RTs shorter than 200 ms were regarded as anticipatory and discarded from the analyses. Additionally, extremely long RT values equivalent to 0.1 % of the RT distribution in each block type (> 2000 ms in pure face

categorization blocks, > 4000 ms in pure voice categorization blocks, and > 7000 ms in mixed blocks) were discarded. Finally, trials involving an error and trials immediately after an error trial were excluded from RT analyses. All results are summarized in Table 1.

Table 1

*Response times (RTs in ms) and error rates (%) as a function of experimental conditions.*

<b>Block Type</b>	<b>Task</b>	<b>Congruency</b>	<b>Sequence Type</b>	<b>RT (SE)</b>	<b>Errors (SE)</b>
Pure	Voice Categorization	Congruent		760 (23.8)	6.9 (0.9)
		Incongruent		810 (24.3)	10.9 (1.2)
	Face Categorization	Congruent		524 (12.6)	4.1 (0.4)
		Incongruent		537 (13.5)	4.6 (0.4)
Mixed	Voice Categorization	Congruent	Repetition	1116 (47.0)	6.3 (0.8)
			Switch	1435 (72.3)	7.8 (1.0)
		Incongruent	Repetition	1170 (41.2)	12.7 (1.4)
			Switch	1496 (59.3)	17.7 (2.4)
	Face Categorization	Congruent	Repetition	995 (48.6)	5.9 (0.9)
			Switch	1294 (65.3)	7.5 (0.9)
		Incongruent	Repetition	1047 (48.8)	13.4 (1.5)
			Switch	1441 (63.9)	18.8 (2.0)

*3.2 Pure block performance.* Results regarding pure blocks are depicted in Figure 1.

The ANOVA regarding RTs revealed a significant main effect of congruency,  $F(1, 23) = 103.82, p < .001, \eta_p^2 = .819$ , indicating overall longer RTs in incongruent conditions (674 ms) than in congruent conditions (642 ms). This is in line with Hypothesis 1. The main effect of task was significant, too,  $F(1, 23) = 185.22, p < .001, \eta_p^2 = .890$ , indicating overall longer RTs for voice (vs. face) categorization (785 ms vs. 531 ms). The interaction was also significant,  $F(1, 23) = 20.30, p < .001, \eta_p^2 = .469$ , reflecting a larger congruency effect for voice (vs. face) categorization (effect of 50 ms vs. 13 ms). This is in line with Hypothesis 2.

Post-hoc contrasts revealed that the congruency effect was significant for both tasks,  $p$ s < .001.

The ANOVA regarding error rates showed a significant main effect of congruency,  $F(1, 23) = 15.54$ ,  $p = .001$ ,  $\eta_p^2 = .403$ , indicating overall higher error rates in incongruent conditions (7.8 %) than in the congruent conditions (5.5 %). Again, this is in line with Hypothesis 1. The main effect of task was significant, too,  $F(1, 23) = 36.81$   $p < .001$ ,  $\eta_p^2 = .615$ , indicating overall more error rates for voice (vs. face) categorization (8.9 % vs. 4.4 %). The interaction was also significant,  $F(1, 23) = 5.44$ ,  $p = .029$ ,  $\eta_p^2 = .191$ , indicating a larger congruency effect for voice (vs. face) categorization (effect of 4.03 % vs. 0.45 %). This is in line with Hypothesis 2. Post-hoc contrasts revealed that the congruency effect was only significant for voice categorization ( $p = .004$ ), not for face categorization ( $p = .341$ ).

To test whether the congruency effect significantly diminished over the course of the pure blocks (time-on-task effects predicted by a statistical learning account), all trials were divided into 4 consecutive quarters. The factor quarter was then additionally entered as a factor into the ANOVA. The crucial interaction of congruency and quarter was far from significant for RTs,  $F(3, 69) = 1.39$ ,  $p = .252$ ,  $\eta_p^2 = .057$ , and error rates,  $F < 1$  (lack of support for Hypothesis 3).

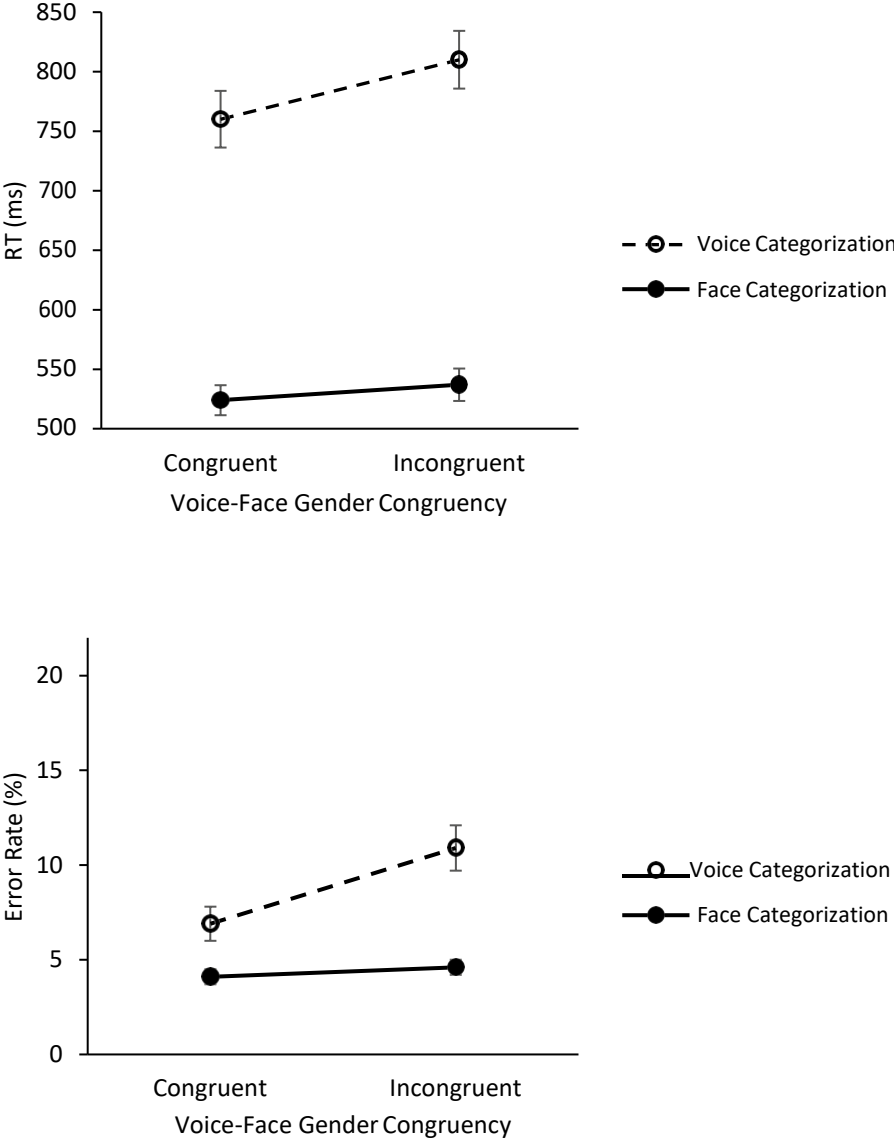


Figure 1. RTs and error rates in pure blocks. Error bars represent standard errors.

3.3 Mixed block performance. Results regarding mixed blocks are depicted in Figure 2. The ANOVA regarding RTs revealed a significant main effect of congruency,  $F(1, 23) = 42.76, p < .001, \eta_p^2 = .650$ , indicating overall longer RTs in incongruent conditions (1289 ms) than in congruent conditions (1210 ms). Again, this is in line with Hypothesis 1. The main effect of task,  $F(1, 23) = 13.77, p = .001, \eta_p^2 = .374$ , indicates overall longer RTs for voice (vs. face) categorization (1304 ms vs. 1194 ms).



There was also a significant main effect of sequence type (switch vs. repetition),  $F(1, 23) = 115.39, p < .001, \eta_p^2 = .834$ , with longer RTs in switch (vs. repetition) trials (1416 ms vs. 1082 ms). The interaction of congruency and task,  $F(1, 23) = 3.85, p = .062, \eta_p^2 = .143$ , was not significant (congruency effect for voice/face categorization: 99/57 ms). However, the significant interaction of congruency and sequence type,  $F(1, 23) = 4.58, p = .043, \eta_p^2 = .166$ , revealed a larger congruency effect in switch (vs. repetition) trials (effect of 103 ms vs. 53 ms). This is evidence in favor of Hypothesis 4a. The interaction of task and sequence type was not significant,  $F < 1$ . In contrast, the three-way interaction was significant,  $F(1, 23) = 8.37, p = .008, \eta_p^2 = .267$ : Interestingly, there was no indication of visual dominance in repetition trials (congruency effect of 54 ms for voice categorization vs. 52 ms for face categorization), but an indication of auditory (voice) dominance (or “reversed” visual dominance) in switch trials (congruency effect of 61 ms for voice categorization vs. 147 ms for face categorization).

The ANOVA regarding error rates showed a significant main effect of congruency,  $F(1, 23) = 64.85, p < .001, \eta_p^2 = .738$ , indicating overall more errors in incongruent conditions (15.7 %) than in congruent conditions (6.9 %). Again, this corroborates Hypothesis 1. There was no significant main effect of task,  $F < 1$ , with error rates of 11% in both task conditions. A significant main effect of sequence type,  $F(1, 23) = 17.21, p < .001, \eta_p^2 = .428$ , revealed more errors in switch (vs. repetition) trials (12.9 % vs. 9.6 %). The interaction of congruency and task,  $F(1, 23) = 1.35, p = .257, \eta_p^2 = .056$ , was not significant. However, the interaction of congruency and sequence type,  $F(1, 23) = 6.39, p = .019, \eta_p^2 = .217$ , revealed significantly greater congruency effects in switch (vs. repetition) trials (effect of 10.7% vs. 6.9 %), mirroring corresponding findings in RTs (in line with Hypothesis 4a). Neither the interaction of task and sequence type nor the three-way interaction were significant,  $F_s < 1$ .

Post-hoc pairwise comparisons between corresponding congruent and incongruent data points (i.e., for each line in Figure 2) for RTs and error rates revealed significant congruency effects throughout (all  $ps < .006$ ), except for one marginally non-significant contrast (voice categorization RTs in switch trials,  $p = .059$ ). Additionally entering trial quarter for mixed blocks as a factor in the analysis (see corresponding test in pure blocks) revealed that the crucial interaction of congruency and quarter was far from significant for RTs,  $F(3, 69) = 1.62, p = .193, \eta_p^2 = .066$ , and error rates,  $F < 1$  (lack of support for Hypothesis 3).

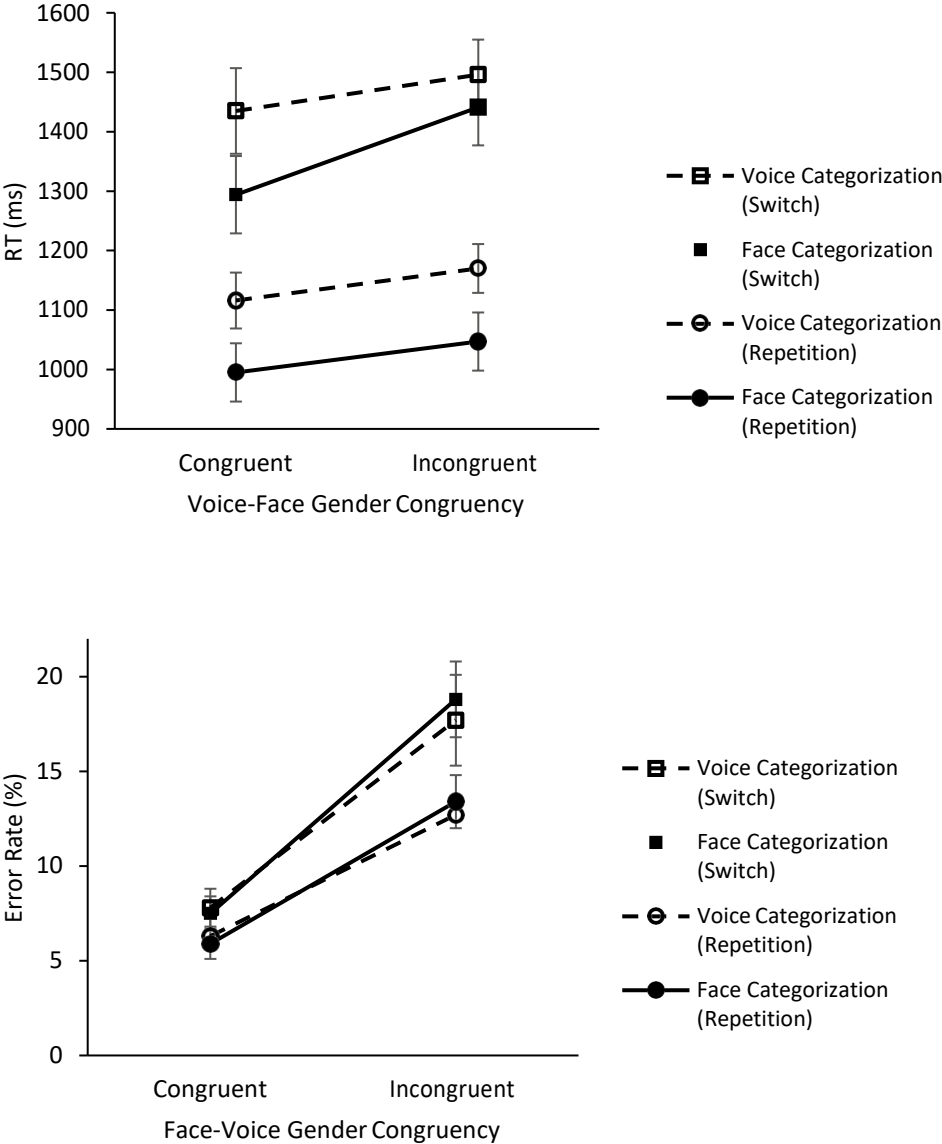


Figure 2. RTs and error rates in mixed blocks. Error bars represent standard errors.

Note that the overall pattern of results (in terms of significant and non-significant effects in RTs and error rates) in both pure and mixed blocks remained the same when discarding all trials involving self-synchronized videos (i.e., videos in which the face and voice belonged to the same person but were synchronized based on different takes, see Section 2.2), suggesting that these self-synchronized videos did not drive the congruency effects. Finally, we also tested whether an analysis of a gender-homogeneous participant group (using the 19 females only) changed the overall result pattern. However, this was not the case (except for the congruency\*task interaction in error rates of pure blocks, which changed from significant to marginally significant). This result shows that the male/female distribution heterogeneity in our sample cannot explain the observed effects.

#### **4. Discussion**

Based on current face-voice processing theories [13], the present study implemented a dynamic face-voice gender interference paradigm to determine the presence and extent of bidirectional integrative interaction between face-based and voice-based gender processing. Participants were presented with short videos of faces with either gender-congruent or gender-incongruent voices and were asked to categorize the gender of either the face or the voice via key press. Specifically, we addressed several hypotheses regarding the presence, strength, generalizability, and determinants of congruency effects indicating interactive face and voice gender processing.

*4.1 Main results and theoretical implications.* Taken together, the present results provide strong evidence in RTs and error rates for congruency effects (Hypothesis 1), indicating a bidirectional interaction of face and voice processing streams regarding gender information. This supports face-voice processing theories assuming a strong interaction of face and voice processing based on a common type of to-be-processed information (here:

gender) [13]. However, the present results suggest that models of voice processing which assume three parallel pathways devoted to speech, emotion, and identity should probably be augmented by a further directed visual processing pathway, in which gender (and age etc.) information can be selectively processed (similar to the original face processing model by Bruce & Young [9], on which current face-voice integration models are based). Previous research regarding gender congruency effects only studied a single direction of interference and focused on utilizing gender-ambiguous stimuli [31, 33, 35], or used static faces which are not suited to address the dynamic process of face-voice integration [36]. Therefore, these previous studies were not able to determine whether face and voice processing is *generally* integrated, since it remained possible that interaction effects might only emerge under specific conditions, for example, when information in one channel does not provide sufficient information to clearly categorize gender.

In contrast, the results of the present study show that congruency effects indeed generalize to non-ambiguous, dynamic conditions typical for daily life situations. The non-ambiguity of our stimuli is confirmed by the fact that mean response accuracy in congruent conditions was very high (consistently above 92%), while the few remaining errors are typical for performance in speeded RT tasks. Additionally, our results show that congruency effects also generalize to situations characterized by high executive control demands (task switching in mixed blocks), and that they do not significantly diminish over the course of the experiment (time-on-task), rendering it unlikely that the congruency effect only reflects a behavioral response to novel (unusual) combinations of gender-incongruent stimulus features especially at the beginning of the respective blocks of trials (as predicted by a statistical learning account [39], see Hypothesis 3). The observation of general congruency effects also speaks against the assumption of strong cognitive modularity in terms of encapsulated processing modules [1], and against the idea that selective attention can perfectly be guided

by input modality [2]. Thus, the ability for selective activation of face or voice processing [3, 11] appears to be severely limited.

Our finding of visual dominance, that is, a stronger effect of response-irrelevant face information on voice processing than vice versa (Hypothesis 2) – which we observed in pure blocks – is in line with similar previous reports of a dominance of face over voice processing. For example, the Facial Overshadowing effect [47] suggests that recognition of a once-heard voice is better when the voice is presented in isolation than with its face [48, 49]. Other research suggests that voices, compared with faces, are a relatively weak cue to a person's identity [11, 50] and to semantics associated with a person [51]. These differences between face and voice processing could only be extinguished when faces are presented in visually degraded fashion [52]. One would probably need independent baseline data for assessing the reliability of faces vs. voices for gender categorization to finally assess whether the visual dominance observed here is due to the participants' greater reliance on visual cues specifically for the gender categorization task, or whether it is indicative of a general, context-independent visual dominance in information processing [35, 36]. Interestingly, the mixed blocks did not reveal evidence for visual dominance, probably suggesting that visual processing prioritization is resource-demanding, so that it is no longer possible when cognitive resources are strongly devoted to executive control demands necessary in mixed task blocks.

Finally, results from the mixed blocks revealed that conditions associated with higher executive control demands (switch trials as opposed to repetition trials) were associated with greater congruency effects (Hypothesis 4a). This could be explained by assuming that under high cognitive load, fewer resources are available to be devoted to gender interference resolution in terms of inhibiting task-irrelevant (distracting) information, eventually resulting in larger interference (congruency) effects. Similar results have also been reported in a

previous study that focused on the impact of working memory load on the ability to inhibit the processing of distracting facial information [53].

*4.2 Limitations and Outlook.* By only implementing congruent and incongruent face-voice stimuli, the present study is unable to finally conclude whether the observed congruency effects are based on facilitation processes (see effects of cross-modal face-voice priming [10, 54]) in the case of congruent stimuli or on interference (e.g., based on inhibition) in the case of incongruent stimuli, or on a mixture of both mechanisms (an equivalent discussion of this theoretical issue has occurred in the context of the Stroop effect [55]). However, creating a good baseline condition is far from trivial. For example, this issue could be addressed in future research involving a condition in which the voice information is accompanied by scrambled visual input merged from both male and female faces preventing visual gender identification (a similar procedure might be implemented for making the voice input unidentifiable in terms of its gender). Such a follow-up study should also, unlike the present study, involve a balanced male/female sample of participants to be able to thoroughly assess potential gender interactions between participants and stimuli.

Furthermore, it appears promising to follow up on the visual dominance effect. Specifically, by manipulating the onset delay between voice and face presentation (e.g., by masking the first frames of the visual input) it could be tested whether and how the visual dominance effect might be attenuated (or even reversed) when access to visual information is delayed [56]. Another interesting field of research would be to study visual dominance in pre-school and school children to address the developmental trajectory of this effect [57, 58].

The congruency effects that emerged in the present study suggest that we never listen to voices or look at faces in isolation, but always attend to the person and its gender as a whole. This may also have implications for the treatment of transgender people, for example, those whose voices are not sufficiently congruent with their visual appearance. Specifically,

our data suggest that further support to increase the matching of visual and voice-related gender information (using voice training, surgery, or hormone intervention) may help to improve communication and overall perceived gender authenticity. The present methodological approach might therefore also be useful in the context of transgender research, for example, by using congruency effects for stimuli involving transgender people as a marker for visual-auditory gender congruency perception.

Finally, it would be interesting to study effects of gender incongruency in tasks where the task-relevant information (i.e., the information that should be responded to) is not related to gender at all. Such a setup would help addressing the question of whether speech information uttered by incongruent, non-authentic speakers (i.e., with voices that violate face-based expectancies) is harder to process compared to speakers whose voices do not violate our expectancies based on visual information.

*4.3 Conclusion.* In the present study, participants constantly had to cross gender borders by switching between male/female categorizations regarding gender-congruent and gender-incongruent face-voice videos. The results indicate the presence of significant congruency effects in both directions – incongruent facial information negatively affected voice gender categorization performance and vice versa. However, the former effect was stronger, supporting theories postulating visual dominance in face-voice integration. Congruency effects were enhanced under high cognitive demands, suggesting the availability of fewer attentional resources for inhibiting distracting information. Overall, we propose that voices are not processed in isolation but in conjunction with facial information, yielding easier processing of more authentic voices, that is, those which do not violate face-based expectancies. The data strengthen theories of face-voice processing which emphasize strong integration between both processing channels.

### **Footnote**

<sup>1</sup>In the present paper, we decided to use the term “gender” when referring to the male/female dichotomy underlying tasks involving the categorization of biological sex. While we are aware that it may be advisable to distinguish between sex as a biologically defined term and gender (identity) as referring to self-definitions of an individual, we decided to align with the terminology used in the majority of previous literature on comparable tasks [22, 30, 31, 32, 34, 35].

### **Acknowledgements**

The authors would like to thank Vera Volk for her help in data acquisition, Stephan Suschke and his Würzburg team of actors at Mainfranken Theater Würzburg for providing stimulus models, and Lynn Huestegge for help with data analyses and for critical feedback on earlier versions of the manuscript.



### References

- [1] Fodor, J. A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.
- [2] Broadbent, D. E. (1958). *Perception and communication*. New York: Oxford University Press.
- [3] Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends Cogn Sci*, 17(6), 263-271.
- [4] Barton, J. J. (2008). Structure and function in acquired prosopagnosia: lessons from a series of 10 patients with brain damage. *J Neuropsychol*, 2, 197-225.
- [5] Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clin Neurophysiol*, 120, 453-463.
- [6] Hailstone, J. C., Crutch, S. J., Vestergaard, M. D., Patterson, R. D., & Warrena, J. D. (2010). Progressive associative phonagnosia: A neuropsychological analysis. *Neuropsychologia*, 48(4), 1104-1114.
- [7] Herald, S. B., Xu, X., Biederman, I., Amir, O., & Shilowich B. E. (2014). Phonagnosia: A voice homologue to prosopagnosia. *Vis Cogn*, 22(8), 1031-1033
- [8] Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: a dissociation between familiar and unfamiliar voices. *Cortex*, 24(2), 195-209.
- [9] Bruce, V., & Young, A. (1986). Understanding face recognition. *Br J Psychol*, 77(3), 305-327.
- [10] Stevenage, S. V., Hugill, A., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *J Cogn Psychol*, 24(4), 409-419.
- [11] Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *Br J Psychol*, 88(1), 143-156.

- [12] Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychol Belg*, *54*, 266-281.
- [13] Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends Cogn Sci*, *11*(12), 535-543.
- [14] Colavita, F. B. (1974). Human sensory dominance. *Percept Psychophys*, *16*, 409-412.
- [15] Spence, C. (2009). Explaining the Colavita visual dominance effect. *Prog Brain Res*, *176*, 245-258.
- [16] Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychon Bull Rev*, *5*, 482-489.
- [17] Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory “compellingness” in the ventriloquism effect: Implications for transitivity among the spatial senses. *Percept Psychophys*, *30*(6), 557-64.
- [18] Shams, L., Kamitani, Y. & Shimojo, S. (2002). Visual illusion induced by sound. *Cogn Brain Res*, *14*, 147-152.
- [19] Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends Cogn Sci*, *8*(3), 129-135.
- [20] Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *Br J Psychol*, *102*, 711-725.
- [21] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- [22] Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Percept Psychophys*, *50*(6), 524-536.

- [23] de Gelder, B. & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cogn Emot*, 14(3), 28-311.
- [24] Hagan, C. C., Woods, W., Johnson, S., Calder, A. J., Green, G. G., & Young A. W. (2009). MEG demonstrates a supra-additive response to facial and vocal emotion in the right superior temporal sulcus. *Proc Natl Acad Sci U S A*, 106(47), 20010-20015.
- [25] Hietanen, J. K., Leppänen, J. M., Illi, M., & Surakka, V. (2004). Evidence for the integration of audiovisual emotional information at the perceptual level of processing. *Eur J Cogn Psychol*, 16(6), 769-790.
- [26] Pourtois, G., de Gelder, B., Bol, A., & Crommelinck, M. (2005). Perception of facial expressions and voices and of their combination in the human brain. *Cortex*, 41(1), 49-59.
- [27] Boltz, M. G. (2017). Facial biases on vocal perception and memory. *Acta Psychol*, 177, 54-68.
- [28] Schweinberger, S. R., Kloth, N., & Robertson, D. M. C. (2011). Hearing facial identities: Brain correlates of face–voice integration in person identification. *Cortex*, 47(9), 1026-1037.
- [29] Zweig, L. J., Suzuki, S., & Grabowecky, M. (2015). Learned face-voice pairings facilitate visual search. *Psychon Bull Rev*, 22(2), 429-436.
- [30] Joassin, F., Maurage, P., & Campanella, S. (2011). The neural network sustaining the crossmodal processing of human gender from faces and voices: an fMRI study. *Neuroimage*, 54(2), 1654-1661.
- [31] Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Curr Biol*, 17(19), 1680-1685.

- [32] Masuda, S., Tsujii, T. & Watanabe S. (2005). An interference effect of voice presentation on face gender discrimination task: Evidence from event-related potentials. *Int Congr Ser*, 1278, 156-159.
- [33] Freeman, J. B., & Ambady, N. (2011). When two become one: Temporally dynamic integration of the face and voice. *J Exp Soc Psychol*, 47, 259-263.
- [34] Weston, P. S., Hunter, M. D., Sokhi, D. S., Wilkinson, I. D., & Woodruff, P. W. (2015). Discrimination of voice gender in the human auditory cortex. *Neuroimage*, 15, 105, 208-14.
- [35] Peynircioglu, Z., Brent, W., Tatz, J., & Wyatt, J. (2017). McGurk effect in gender identification: Vision trumps audition in voice judgments. *J Gen Psychol*, 144(1), 59-68.
- [36] Latinus, M., VanRullen, R., & Taylor, M. J. (2010). Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC Neurosci*, 11: 36.
- [37] Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J Exp Psychol*, 18(6), 643-662.
- [38] MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychol Bull*, 109, 163-203.
- [39] Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends Cogn Sci*, 19, 117-125.
- [40] Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance series. Attention and performance 15: Conscious and nonconscious information processing* (pp. 421-452). Cambridge, MA: MIT Press.

- [41] Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *J Exp Psychol Gen*, *124*(2), 207-231.
- [42] Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychol Bull*, *136*(5), 849-874.
- [43] Alsius A., Möttönen R., Sams M. E., Soto-Faraco S. & Tiippana K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Front Psychol*, *15*(5), 727.
- [44] Kim, S. Y., Kim, M. S., & Chun, M. M. (2005). Concurrent working memory load can reduce distraction. *Proc Natl Acad Sci U S A*, *102*, 16524-9.
- [45] Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Mem Cognit*, *3*, 166-174.
- [46] Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *J Exp Psychol Learn Mem Cogn*, *22*, 1423-1442.
- [47] Cook, S., & Wilding, J. (1997). Earwitness Testimony 2: Voices, Faces and Context. *Appl Cogn Psychol*, *11*(6), 527-541.
- [48] Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Appl Cogn Psychol*, *25*(1), 112-118.
- [49] Stevenage, S. V., Neil, G. J., & Hamlin I. (2014). When the face fits: recognition of celebrities from matching and mismatching faces and voices. *Memory*, *22*(3), 284-94.
- [50] Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *Q J Exp Psychol*, *51*(1), 179-195.

- [51] Barsics, C., & Brédart, S. (2012a). Recalling semantic information about newly learned faces and voices. *Memory*, *20*(5), 527-534.
- [52] Hanley, J. R., & Damjanovic, L. (2009). It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. *Memory*, *17*(8), 830-9.
- [53] De Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science*, *291*, 1803-1806.
- [54] Schweinberger, S. R., Robertson, D., & Kaufmann, J. M. (2007). Hearing facial identities. *Q J Exp Psychol*, *60*, 1446-1456.
- [55] Dyer, E. N. (1973). The Stroop phenomenon and its use in the study of perceptual, cognitive, and response processes. *Mem Cognit*, *1*, 106-120.
- [56] Robertson, D. M. C., & Schweinberger, S. R. (2010). The role of audiovisual asynchrony in person recognition. *Q J Exp Psychol*, *63*(1), 23-30.
- [57] Nava E. & Pavani F. (2013). Changes in sensory dominance during childhood: converging evidence from the colavita effect and the sound-induced flash illusion. *Child Dev*, *84*(2), 604-16.
- [58] Noles, N. S., & Gelman, S. A. (2012). Preschool children and adults flexibly shift their preferences for auditory versus visual modalities, but do not exhibit auditory dominance. *J Exp Child Psychol*, *112*(3), 338–350.

# Study 1

Huestegge, S., & Raettig, T. Crossing gender borders: Bidirectional dynamic interaction between face-based and voice-based gender categorization

In press: *Journal of Voice* (accepted for publication on 25.09.2018)  
Published "online first": <https://doi.org/10.1016/j.jvoice.2018.09.020>

Bei Studie 1 handelt es sich um eine Version (post-review, pre-proof) des o.g. Artikels (Nutzung hier mit freundlicher Genehmigung des Elsevier-Verlags).

# Study 2

Huestegge, S. M., Raettig, T., & Huestegge, L. (2019). Are face-incongruent voices harder to process? Effects of face-voice gender incongruency on basic cognitive information processing. *Experimental Psychology*, 66, 154-164. (accepted for publication on 10.01.2019)

Bei Studie 2 handelt es sich um eine Version (post-review, pre-proof) des o.g. Artikels (Nutzung hier mit freundlicher Genehmigung von *Experimental Psychology* 2019; Vol. 66(2), 154–164 ©2019 Hogrefe Publishing; [www.hogrefe.com](http://www.hogrefe.com)).

DOI: <https://doi.org/10.1027/1618-3169/a000440>

# Study 3

Sujata M. Huestegge (subm.).. Matching unfamiliar voices to static and dynamic faces: No evidence for a dynamic face advantage in a simultaneous presentation paradigm

Submitted: *Frontiers: Psychology* (15.02.2019)

Bei Studie 3 handelt es sich hier um eine Version (initially submitted version, but without any changes that occurred during the revision process) des o. g. Artikels.

Die finale Version wurde am 08.08.2019 von *Frontiers: Psychology* akzeptiert und kann unter DOI: 10.3389/fpsyg.2019.01957 (Copyright © 2019 Huestegge) abgerufen werden.

**Are face-incongruent voices harder to process?**

**Effects of face-voice gender incongruency on basic cognitive information processing**

Sujata M. Huestegge<sup>1, 2</sup>, Tim Raettig<sup>1</sup>, & Lynn Huestegge<sup>1</sup>

<sup>1</sup>University of Würzburg

<sup>2</sup>University of Music and Performing Arts Munich

Word count: 5500 (text body only, without references etc.)

Correspondence address:

Sujata M. Huestegge

University of Music and Performing Arts Munich

Arcisstr. 12

80333 Munich, Germany

sujata.huestegge@hmtm.de



### **Abstract**

Based on current integration theories of face-voice processing, the present study had participants process 1152 videos of faces uttering digits. Half of the videos contained face-voice gender-incongruent stimuli (vs. congruent stimuli in the other half). Participants indicated digit magnitude or parity. Tasks were presented in pure blocks (only one task) and in task switching blocks (using colored cues to specify task). The results indicate significant congruency effects in pure blocks, but partially reversed congruency effects in task switching, probably due to enhanced assignment of capacity towards resolving difficult situational demands. Congruency effects did not dissipate over time, ruling out that initial surprise associated with incongruent stimuli drove the effects. The results show that interference between two task-irrelevant person-related dimensions (face/voice gender) can affect processing of a third, task-relevant dimension (digit identity), suggesting greater processing ease associated with more authentic voices (i.e., voices that do not violate face-based expectancies).

**Keywords:** Voice Gender Processing; Crosstalk; Face-voice integration; Gender Identity; Visual dominance; Cognitive load; Auditory-visual interference; Task switching; Executive Control

Voices and faces convey a vast variety of information, such as a person's identity, age, affective state, personality, heritage, cultural background, utterances, as well as gender (see Belin, Fecteau, & Bédard, 2004). Based on the observation that similar types of information are processed in both the visual and the auditory channel, combined theories of face and voice processing assume cross-modal interaction, ultimately resulting in reliable information regarding another person's speech content, affect, and identity etc. (Campanella & Belin, 2007; Yovel & Belin, 2013). The present study focuses on potential effects of cross-modal (audio-visual) interactions regarding gender processing on gender-unrelated basic information processing (spoken digit classification).

*Face-voice interaction.* The most famous phenomenon in the context of the effects of face-voice interaction on speech-related information processing is the McGurk effect. Specifically, McGurk and MacDonald (1976) provided substantial evidence for a major influence of vision on auditory speech perception. The study involved stimulus material showing a woman's face while repeating the syllables /ba-ba/, /ga-ga/, /pa-pa/, or /ka-ka/. The auditory and visual tracks of the videos were dubbed in the following four combinations: /ba/-voice + /ga/-lips; /ga/-voice + /ba/-lips, /pa/-voice + /ka/-lips, /ka/-voice + /pa/-lips. Participants (adults, pre-school children, and school children) were asked to repeat what they heard the model saying. A central finding was that especially in adults (in 98% of responses), the /ba/-voice + /ga/-lips combination yielded a (voiced) /da/-response, that is, a fused speech percept. A similar, but attenuated effect was observed in the /pa/-voice + /ka/-lips condition (report of the voiceless /ta/ in 81% of responses in adults). Note that both conditions involved a bilabial plosive in the auditory track that was not mirrored by a corresponding closing of lips in the visual track, therefore producing a need for perceptual disambiguation. In sum, the McGurk effect is the result of audiovisual integration, resulting in a specific auditory illusory percept that was neither present in the visual nor in the auditory channel.

Importantly, Green et al. (1991) demonstrated that the McGurk effect can still be observed even when the presented gender of the heard voice and the visual face did not match, so that both streams of information clearly came from different speakers. Therefore, audiovisual fusion does not necessitate the perception of an integrated source (a single person) for both (auditory and visual) streams of information. The study also suggests that face-voice gender incongruency does *not* affect speech perception (in terms of the McGurk effect).

Another type of gender congruency effects on information processing performance stems from the domain of memory research. It was suggested that information processing in terms of memory retrieval is easier for words when speaker gender during test is identical to speaker gender during encoding (Goldinger, 1996, but see Palmieri, Goldinger, & Pisoni, 1993 and Campeanu, Craik, & Alain, 2013, who found an advantage for identical speakers, not for merely gender-congruent speakers). However, in these studies gender congruency was not manipulated for two simultaneous streams of information but rather across two distinct points in time (encoding and retrieval), and the results are rather ambiguous regarding the extent to which speech-related and (task-irrelevant) voice feature-related processing interacts.

*Speech-voice interaction.* Other experiments further addressed the interdependence of speech processing and voice processing. For example, listeners are able to recognize familiar voices from variable utterances, generally suggesting the presence of *speech-independent* voice representations (Zäske, Volberg, Kovács, & Schweinberger, 2014). In contrast, other phenomena suggest some level of *integration* between voice and speech processing. For example, voice familiarity can increase speech intelligibility (Nygaard, Sommers, & Pisoni, 1994) and speech-based memory performance (Geiselman & Crawley, 1983; Nygaard & Pisoni, 1998). Conversely, speech-related (linguistic) proficiency can be beneficial for voice identification (Perrachione & Wong, 2007). Regarding gender-related effects, it has been shown that voice gender after-effects (i.e., adaptation to male voices causing subsequent

voices to sound more female and vice versa) can occur even when prior attention was directed towards phonetic content instead of voice gender (Zäske, Fritz, & Schweinberger, 2013), suggesting relatively automatic processing of voice gender along with linguistic information.

*Face-voice interaction and gender classification.* Finally, another line of research regarding face-voice interaction is specifically focused on gender information processing by manipulating the congruency of the gender of simultaneously presented voices and faces. Thereby, it is possible to directly assess the effects of cross-modal gender congruency on gender classification in either the visual or auditory modality. Most gender classification studies focus on visual face gender classification while simultaneously presenting gender-congruent or incongruent information in the auditory channel. For example, Smith, Grabowecky, and Suzuki (2007) showed that androgynous faces were more likely to be categorized as female when presented with pure high pitch tones reminiscent of the female fundamental speaking frequency range (vice versa for low pitch tones associated with the male range). In an event-related potential study, Masuda, Tsujii, and Watanabe (2005) measured the N170, a component known to be tightly linked to face processing, and demonstrated that gender-incongruent (as opposed to congruent) voices can modulate this particular component, likely representing a disruption of facial gender processing. Another study employed morphed faces to increase visual gender ambiguity (Freeman & Ambady, 2011). Simultaneously, they presented either original male/female (sex-typical) voices or formant-altered (sex-atypical) voices. Using a mouse tracking paradigm in which gender classification is made by moving a mouse cursor from a starting point to either one of two relatively distant target areas (representing male/female gender), they showed that voice information biased the participants' spatial response trajectory: For example, a sex-atypical voice attenuated the corresponding face gender categorization process.

Until now, studies directly addressing the impact of additional visual face gender information on voice gender discrimination are rare, and either did not assess gender classification directly or did not use dynamic face video material. For example, Peynircioglu, Brent, Tatz and Wyatt (2017) focused on a closely related issue by using brief male/female voice stimuli (pitch was held constant). These auditory stimuli were visually time-synchronized with videos of either a male or a female singer. Subsequent classification performance (which only indirectly assessed gender by asking for voice type classification such as baritone or alto) was biased towards the simultaneously presented visual gender information. Finally, Latinus, VanRullen, and Taylor (2010) studied mutual effects of gender-(in)congruent static faces on voice gender classification and vice versa, showing congruency effects in both directions. However, due to the static face stimuli they did not address face-voice *integration*, which is conceptualized as a dynamic interactive process (Campanella & Belin, 2007). Only recently, work from our own lab for the first time demonstrated bidirectional gender congruency effects on gender classification by using dynamic, time-synchronized videos to address actual integration processes (Huestegge & Raettig, 2018, see Schweinberger, Robertson, & Kaufmann, 2007, for similar effects of audio-visual integration effects on recognition performance instead of gender classification).

Taken together, these previous gender classification studies focused on conflicts between face and voice gender processing while one of these two dimensions was task-relevant. However, none of these previous studies addressed the related, but novel issue whether face-voice gender congruency might affect the processing of information unrelated to gender. This is also important for current theorizing on face-voice processing, for example, because the presence of distinct processing routes for gender (as well as age etc.) information is usually disregarded, and it has been assumed that visual and auditory processing pathways can also be activated selectively (Ellis, Jones, & Mosdell, 1997; Yovel & Belin, 2013). Additionally, it is theoretically interesting whether face-voice gender congruency is processed

relatively automatic, that is, even without direction of attention towards gender via task instructions/requirements (see Discussion for details).

*The present study.* In the present study, we addressed these novel issues by having participants respond to digits uttered in videos depicting either face-voice congruent or incongruent speakers. Two gender-unrelated tasks were used that often serve as typical proxies for basic general information processing (e.g., Kiesel et al., 2010): Participants were either asked to indicate whether a digit is smaller/larger than 5 (magnitude task) or odd/even (parity task).

*1. Effects of face-voice congruency on basic information processing.* On a theoretical level, two outcomes are conceivable. First, one could expect no congruency effect whatsoever, because unlike in the previous (cross-modal conflict) gender classification studies referred to above, and unlike in typical cognitive conflict tasks (such as Stroop tasks, Stroop, 1935, or Flanker tasks, Eriksen & Eriksen, 1974) involving dimensional overlap between stimulus dimensions (see Kornblum, Hasbrouck, & Osman, 1990), the present setup involves conflict between two task-irrelevant stimulus dimensions. Furthermore, there is neither dimensional overlap between the task-relevant and the task-irrelevant stimulus dimensions, nor between the stimulus dimensions associated with gender congruency and the (left/right) responses. However, classic theories of conflict resolution assume that these types of dimensional overlap are necessary to observe interference effects in the first place (Kornblum et al., 1990). Instead, this viewpoint would predict that selective attention should work rather flawlessly (in terms of focusing solely on the relevant digit information), enabling participants to disregard potential conflicts between the task-irrelevant stimulus dimensions. In the case of face-voice integration, this view appear to be supported by the study by Green et al. (1991, see above), since it suggests that face-voice gender incongruency may *not* affect speech perception.

In contrast to such an abstract view of information processing in terms of (objectively defined) task-relevant processing codes and associated dimensional overlap of code sets, recent years have witnessed growing support for an embodied cognition perspective (e.g., Barsalou, 2008, Barsalou, Simmons, Barbey, & Wilson, 2003). One central tenet of this view is that mental representations are grounded in (or represented in terms of) their sensory (or motor) origin and its situational context (situated cognition, see Wilson, 2002). When applied to the present setup, this would imply that representations of digits still carry with them the context of their sensory origin, including the (task-irrelevant) facial and vocal features and their interrelation (i.e., congruency), ultimately being able to affect cognitive performance. Thus, this view would predict the emergence of congruency effects despite the fact that these are not dimensionally related to the instructed task requirements, and some previous studies on voice processing already suggest that voice features (including gender) may indeed be encoded relatively automatically (e.g., Zäske et al., 2013).

2. *Congruency and time-on-task.* If a congruency effect is found in the two tasks, it is still possible that it is merely due to initial surprise associated with the experience of artificial (non-authentic) face-voice gender incongruent stimuli at the beginning of the experiment, rather than representing evidence for a generic inability to disregard task-irrelevant speaker information for information processing. Also, familiarity with the (initially unfamiliar) stimulus material should increase over the course of the experiment, a factor known to potentially affect speech processing (e.g., Schweinberger & Robertson, 2017). This was addressed by testing for differences between congruency effects (if present) over the course of the respective task blocks (time-on-task).

3. *Impact of executive control demands.* Finally, we were interested in testing whether general executive control demands modulate the congruency effect (if present). To manipulate executive cognitive control demands, tasks were not only presented in pure blocks (only one

task per block of trials), but also in task switching blocks (using colored cues to specify the task in each trial, see Meiran, 1996). Task switching (Allport, Styles, & Hsieh, 1994, Rogers & Monsell, 1995) is assumed to be associated with mental task set reconfiguration and interference between a current task set and an alternative previous task set, thereby substantially increasing mental load associated with these executive control demands (see Kiesel et al., 2010). On the basis of previous research on mental load and basic cognitive processes (Kim, Kim, & Chun, 2005; Murphy, Groeger, Greene, 2016), two potential outcomes appear principally conceivable.

First, one could expect *stronger* congruency effects in switch (vs. repetition) conditions, because higher cognitive demands necessary to complete a switch trial might consume resources needed for resolving interference based on gender incongruency (see Murphy et al., 2016, for a review of similar cognitive load effects on performance). Furthermore, cognitive load (in terms of working memory load) has already been shown to increase effects related to task-irrelevant voice processing (voice-gender after-effects; Zäske, Perlich, & Schweinberger, 2016). Similarly, in a previous study we showed that face-voice congruency effects on gender categorization were indeed larger under strong executive control demands (task switching vs. task repetition, Huestegge & Raettig, 2018).

Second, it is also possible to expect *attenuated* congruency effects in switch (vs. repetition) trials, since the increased consumption of cognitive resources for the completion of switch trials may yield a decrease of resources available for the processing of task-irrelevant information (i.e., even fewer attentional focus on information which is not relevant for the categorization task, see Logan & Zbrodoff, 1979). This is especially likely because the task at hand (digit classification) does not require gender interference resolution in the first place, so that an increase in cognitive load should attenuate (or even extinguish) congruency effects



(see also Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014, for evidence of a weaker McGurk effect under high cognitive load).

### Method

*Participants.* Twenty-four participants took part in the experiment (2 male, mean age = 22 years,  $SD = 3.35$ , range = 18-31). A power analysis based on related data of another study in which we found strong gender congruency effects on gender classification indicated that fewer than 5 participants were needed to achieve a power of  $1-\beta = .95$ . Due to the novel focus on effects of congruency between task-irrelevant stimulus dimensions, we opted for a substantially larger sample size. All participants had normal (self-reported) hearing and normal or corrected-to-normal vision. They received monetary reimbursement (or a small present) for participation. All participants gave informed consent.

*Stimuli & Apparatus.* Participants were seated about 67 cm in front of a TFT computer screen (15'') with a standard computer keyboard in front of them. Two keyboard keys (left Ctrl, Alt) served as response keys for the magnitude and parity tasks. Stimuli (available upon request) were short (fixed duration of 1 s; 25 frames) video clips of male or female faces uttering digits (1, 2, 3, 4, 6, 7, 8 or 9) with a time-synchronized (i.e., A-V synchronized) male or female voice (presented via headphones). The stimuli were based on original video recordings of three male and three female speakers (professional actors, unfamiliar to participants) uttering the digits. Since we aimed at natural visual face presentation with clear gender cues, the videos also showed the hair and upper torso of the actors. The six actors (age range: 32-53 years, dialect-free native speakers, neutral facial expression) were selected to ensure that both their visual appearance as well as their voice was unambiguously male or female. This was validated by informally presenting either the faces or the voices (uttering the digits) to 12 participants (no overlap with the sample from the main experiment), who correctly (100%) categorized gender of all six faces and voices. Male and female actors were

(within reasonable limits) matched regarding their age. To rule out any potential advantage associated with unsynchronized (original) videos (which could only occur in the congruent condition), all videos were synchronized (using Pinnacle 21 software), that is, for all stimuli the heard voice was never that of the original video. However, the stimuli also included self-synchronized videos, that is, videos in which the voice sample from one take of a participant uttering a digit was synchronized with the visual sample of another take from the same participant. It was ensured that voice information (equivalent to the visual face information) was immediately present in the first frame. A visual fade out was implemented at the end of each video (final 8 frames) to ensure a smooth visual experience. In half of the stimuli the gender of the voice matched with that of the face, whereas the other half consisted of incongruent face-voice pairings. Altogether, these voice-face pairings resulted in  $8 \text{ (digits)} * 36 \text{ (voice-face pairings)} = 288$  stimuli (half of them congruent vs. incongruent regarding face/voice gender) that were presented in the middle of the screen ( $22^\circ * 12^\circ$  visual angle) on black background.

*2.3 Procedure.* The experiment (using the software Presentation®, Version 19.0, Neurobehavioral Systems, Inc., Berkeley, CA) started with an instruction screen. Participants in the pure blocks (i.e., blocks with only one task) were asked to attend to the videos (preceded by a central fixation cross presented for 250 ms) and to respond as fast and accurately as possible by pressing the left key for either “< 5” or “odd” and the right key for either “> 5” or “even” in the magnitude or parity task, respectively. They were instructed to look at the videos throughout. Task order of the pure blocks (magnitude/parity task) was counterbalanced across participants. Each of the 288 stimuli was shown twice during each half of the pure blocks (i.e., for each of the two tasks). Additionally, participants completed two task switching blocks (also involving  $2*288$  stimuli) consisting of random trial-by-trial switches (switching probability: 50%) between the magnitude and the parity task. The required task in each trial was indicated by a red vs. green rectangle (cued task switching, see

Meiran, 1996) surrounding the face video (color-task assignment counterbalanced across participants). Note that the pure blocks also contained these rectangles of different color in random order, but participants were instructed to ignore them. Block order (pure blocks first vs. task switching blocks first) was counterbalanced across participants. All stimuli were presented in random order within each block. The inter-trial interval was constant and relatively short (100 ms). The experiment lasted about 40 minutes.

*2.4 Design.* Reaction times (RTs) and error rates (%) served as dependent variables. Separate ANOVAs for each dependent variable and for each block type (pure vs. task switching) were computed. For pure blocks, the analysis involved the within-subject factors congruency (congruent vs. incongruent face-voice pairings) and task (magnitude vs. parity task). For task switching blocks, the analysis additionally involved the factor sequence type (switch vs. repetition, i.e., whether the trial represented a task switch or repetition relative to the previous trial). To test whether the gender congruency effect, if present, is affected by time-on-task, we additionally divided trials (of pure blocks and task switching blocks) into four consecutive quarters and tested for significant interactions of congruency with quarter as an additional factor. In an additional analysis, we also focused on cross-task response congruency as another factor associated with cognitive load (by comparing trials involving digits that were mapped to the same vs. different response keys across the two tasks: 1, 3, 6, 8 vs. 2, 4, 7, 9), to test for modulations of congruency effects. This variable reflects the extent to which the currently irrelevant task (specifically its associated responses) interferes with the currently relevant task.

## Results

*Data treatment.* For the RT analyses, RTs shorter than 150 ms were regarded as anticipatory and discarded from the analyses. Additionally, extremely long RT values equivalent to 0.1 % of the RT distribution in each block type (> 3000 ms in pure task blocks

and > 5000 ms in task switching blocks) were discarded. Finally, trials involving an error and trials immediately after an error trial were excluded from RT analyses (yielding 88.21/82.94% valid trials of trials in pure/mixed blocks for RT analyses). The raw data are openly available (URL: <https://doi.org/10.5281/zenodo.2536051>).

*Pure block performance.* Results regarding pure blocks are depicted in Figure 1. The ANOVA regarding RTs revealed a significant main effect of task,  $F(1, 23) = 46.70, p < .001, \eta_p^2 = .670$ , indicating overall longer RTs for the parity (vs. magnitude) task (770 ms vs. 688 ms). The main effect of congruency was significant, too,  $F(1, 23) = 10.31, p = .004, \eta_p^2 = .310$ , indicating overall longer RTs in incongruent conditions (736 ms) than in congruent conditions (722 ms). There was no significant interaction,  $F(1, 23) = 1.28, p = .270, \eta_p^2 = .053$ .

The corresponding ANOVA regarding error rates showed a significant main effect of task,  $F(1, 23) = 17.64, p < .001, \eta_p^2 = .434$ , indicating higher error rates for the parity (vs. magnitude) task (7.2 % vs. 4.7 %). There was neither a significant main effect of congruency, nor a significant interaction,  $F_s < 1$ .

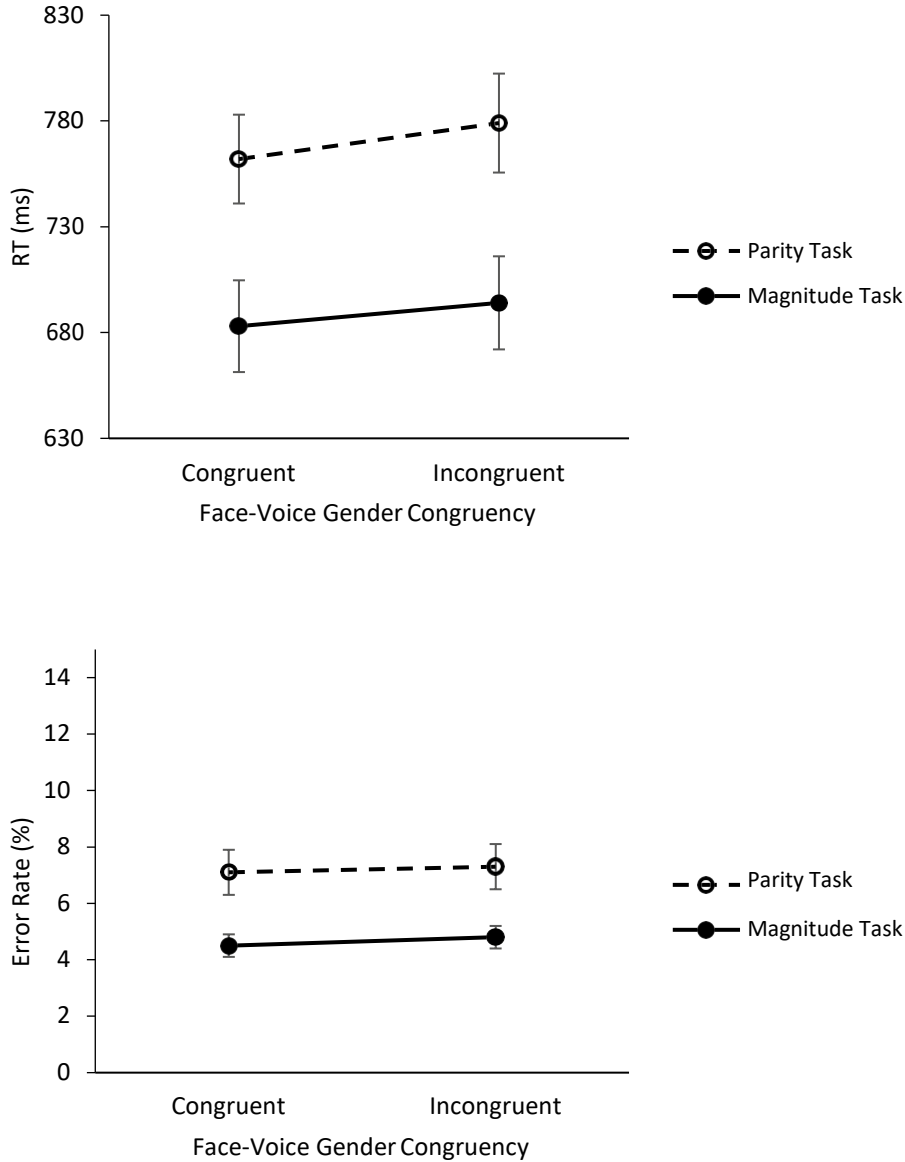


Figure 1. Mean RTs (top) and mean error rates (bottom) in pure blocks. Error bars represent standard errors.

To test whether the congruency effect was modulated by time-on-task, the factor quarter was additionally entered as a factor into the ANOVA. The crucial interaction of congruency and quarter was far from significant for both RTs and error rates,  $F_s < 1$ .

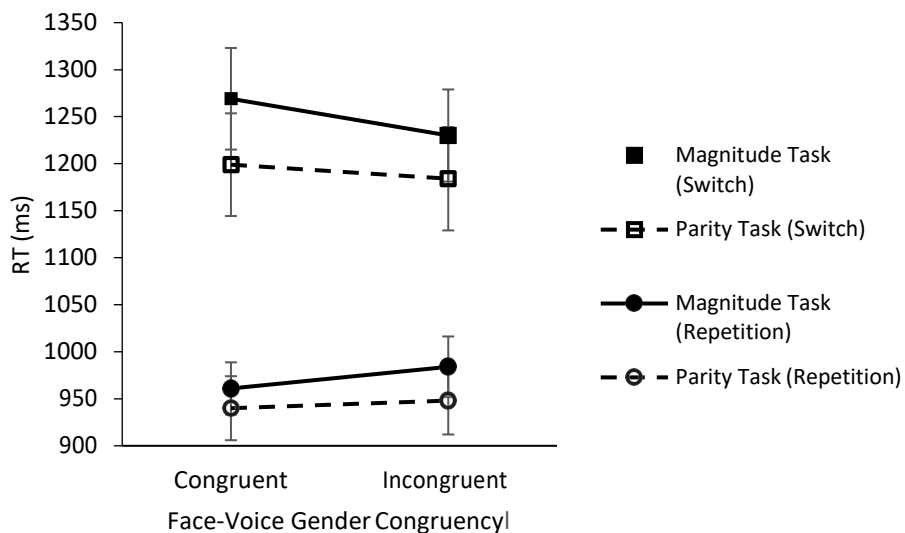
As a final analysis, we focused on potential effects of cross-task response congruency by adding this factor (instead of “quarter”) to the ANOVA. For pure blocks, this analysis only

included participants that started with a switch block (only these participants had experience with the respective other task before experiencing pure blocks to allow for potential interference). There was a significant main effect of cross-task response congruency in RTs,  $F(1, 7) = 19.71, p = .003, \eta_p^2 = .733$  (738 ms in congruent trials, 777 ms in incongruent trials). However, there was no significant interaction of cross-task response congruency and (face-voice) congruency,  $F < 1$ . The same ANOVA for error rates showed no significant main effect of cross-task response congruency,  $F(1, 7) = 4.12, p = .082, \eta_p^2 = .370$  (4.9 % in congruent trials, 7.1 % in incongruent trials), and no significant interaction between cross-task response congruency and (face-voice) congruency,  $F < 1$ , but a significant interaction of cross-task response congruency and task,  $F(1, 7) = 6.24, p = .041, \eta_p^2 = .471$  (congruency effects: 1% in magnitude task, 3.4% in parity task). All other interactions with cross-task response congruency were not significant.

*Task switching block performance.* Results regarding mixed blocks are depicted in Figure 2. The ANOVA regarding RTs showed a significant main effect of task,  $F(1, 23) = 6.20, p = .020, \eta_p^2 = .212$ , indicating overall longer RTs for the magnitude (vs. parity) task (1111 ms vs. 1068 ms). There was no significant main effect of congruency,  $F < 1$ , but a significant main effect of sequence type (switch: 1220 ms vs. repetition: 958 ms),  $F(1, 23) = 62.89, p < .001, \eta_p^2 = .732$ , indicating significant switch costs. The interactions of congruency and task, and of sequence type and task, were not significant,  $F_s < 1$ . Interestingly, there was a significant interaction of congruency and sequence type,  $F(1, 23) = 9.94, p = .004, \eta_p^2 = .302$ , indicating a *reversed* congruency effect in switch vs. repetition trials (-27 ms vs. 16 ms). The three-way interaction was not significant,  $F(1, 23) = 2.29, p = .144, \eta_p^2 = .090$ . To follow up on the interaction indicating a reversed congruency effect, we computed post-hoc contrasts comparing corresponding congruent and incongruent data points (i.e., for each line in Figure 2, without correction for multiple testing). There was no indication for a significant congruency effect in repetition trials (magnitude task:  $p = .112$ , parity task:  $p = .391$ ), but one

significant reversed congruency effect in switch trials, namely in the magnitude task ( $p = .030$ , parity task:  $p = .423$ ).

The corresponding ANOVA regarding error rates showed a significant main effect of task,  $F(1, 23) = 7.67$ ,  $p = .011$ ,  $\eta_p^2 = .250$ , indicating overall more errors in the magnitude (vs. parity) task (10.0 % vs. 8.1 %). There was no significant main effect of congruency,  $F = 1.00$ , but of sequence type,  $F(1, 23) = 26.30$ ,  $p < .001$ ,  $\eta_p^2 = .533$ , revealing more errors in switch (vs. repetition) trials (10.9 % vs. 7.2 %). There was no significant interaction of task and congruency,  $F(1, 23) = 1.65$ ,  $p = .212$ ,  $\eta_p^2 = .067$ . However, a significant interaction of task and sequence type,  $F(1, 23) = 7.64$ ,  $p = .011$ ,  $\eta_p^2 = .249$ , indicated greater switch costs in the magnitude task (5.1 %) than in the parity task (2.1 %). Neither the interaction of task and condition, nor the interaction between congruency and sequence type, nor the three-way interaction were significant,  $F_s < 1$ .



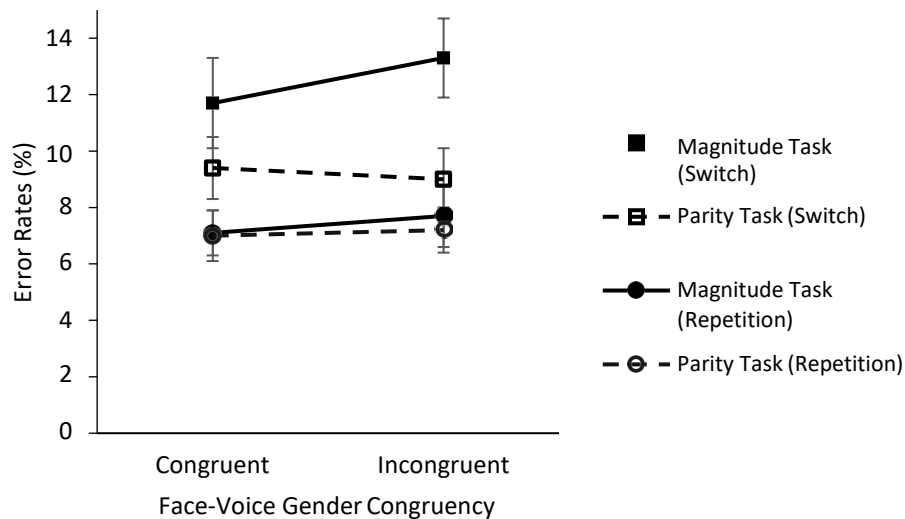


Figure 2. Mean RTs (top) and mean error rates (bottom) in task switching blocks. Error bars represent standard errors.

Additionally adding quarter (of trials in task switching blocks) as a factor to the design revealed no significant interaction of congruency and quarter in RTs,  $F(3, 69) = 1.09$ ,  $p = .362$ ,  $\eta_p^2 = .045$ , or error rates,  $F < 1$ . There was also no significant interaction of congruency, sequence type, and quarter regarding RTs,  $F < 1$ .

We finally focused on effects of cross-task response congruency by adding this factor (instead of quarter) to the ANOVA. There was a significant main effect of cross-task response congruency in RTs,  $F(1, 23) = 64.90$ ,  $p < .001$ ,  $\eta_p^2 = .738$  (1051 ms in congruent trials, 1134 ms in incongruent trials). However, there was no significant interaction of cross-task response congruency and (face-voice) congruency,  $F < 1$ . Cross-task response congruency significantly interacted with sequence type,  $F(1, 23) = 21.80$ ,  $p < .001$ ,  $\eta_p^2 = .487$  (cross-task response congruency effect: 59 ms for repetition trials and 105 ms for switch trials), and with task,  $F(1, 23) = 9.58$ ,  $p = .005$ ,  $\eta_p^2 = .294$  (cross-task response congruency effect: 111 ms for the magnitude task and 53 ms for the parity task). The same ANOVA for error rates showed the



same pattern of results: There was a significant main effect of cross-task response congruency,  $F(1, 23) = 110.70, p < .001, \eta_p^2 = .828$  (5.0 % in congruent trials, 13.1 % in incongruent trials), but no significant interaction between cross-task response congruency and (face-voice) congruency,  $F < 1$ . However, cross-task response congruency significantly interacted with sequence type,  $F(1, 23) = 12.72, p = .002, \eta_p^2 = .556$  (cross-task response congruency effect: 6.2 % for repetition trials and 9.9 % for switch trials), and with task,  $F(1, 23) = 11.71, p = .002, \eta_p^2 = .337$  (cross-task response congruency effect: 11.1 % for the magnitude task and 5.2 % for the parity task). All other interactions with cross-task response congruency were not significant.

Note that the overall pattern of results (in terms of significant/non-significant congruency main effects and interactions in RTs and error rates) in both pure and mixed blocks remained the same when discarding all trials involving self-synchronized videos (i.e., videos in which the face and voice belonged to the same person but were synchronized based on different takes, see method section), suggesting that these self-synchronized videos did not drive the congruency effects.

### Discussion

The present study aimed to test whether face-voice gender incongruency affects basic information processing associated with spoken digits, that is, in a task in which the task-relevant stimulus dimension is not associated with gender information at all. The RT results from the pure blocks unequivocally supported this assumption: Digit processing in both the parity task and the magnitude task was faster when the gender of the voice uttering the digits corresponded to the gender of the face, and this effect was not significantly modulated by time-on-task (ruling out substantial effects of surprise or growing stimulus familiarity). Moreover, the interpretation of this effect was not compromised (e.g., in terms of a speed-accuracy trade-off) by any reversed effect in the error data. Thus, congruency between the

two task-irrelevant stimulus dimensions (voice and face gender) affected the processing of a third, task-relevant stimulus dimension (spoken digit categorisation).

On a general theoretical level, this observation of gender-congruency effects on digit processing speaks against an abstract view of information processing according to which task-relevant processing codes and associated dimensional overlap of code sets essentially determine performance (e.g., Kornblum et al., 1990). Note that previous (cross-modal conflict) gender classification studies (Freeman & Ambady, 2010; Latinus et al., 2010; Masuda et al., 2005, Peynircioglu et al., 2017; Smith et al., 2007), and typical cognitive conflict tasks (such as Stroop or Flanker tasks, Stroop, 1935; Eriksen & Eriksen, 1974) do involve dimensional overlap between stimulus dimensions. Instead, the present results are more compatible with an embodied cognition view (e.g., Barsalou, 2008; Barsalou et al., 2003; Wilson, 2002). Specifically, it appears that mental representations of digits in our study are still associated with the context of their perceptual origin, including the (task-irrelevant) facial and vocal features and their interrelation (i.e., congruency). In this way, speaker gender congruency can affect performance in that non-authentic voices (i.e., those which are unexpected based on facial cues) are harder to process than authentic voices, supporting the assumption that mental representations are grounded in their sensory origin and its situational context (situated cognition, e.g., Greeno, 1998). On a more specific level, these results are also in line with voice-related studies suggesting relatively automatic integration between speech and voice processing (Geiselman & Crawley, 1983; Nygaard et al., 1994; Nygaard & Pisoni, 1998; Perrachione & Wong, 2007; Zäske et al., 2013).

In the context of face-based and voice-based cognition, our results support theories of face and voice processing which assume strong cross-modal interaction for different types of to-be processed information (e.g., Campanella & Belin, 2007; Yovel & Belin, 2013), ultimately resulting in reliable information regarding another person's affect, identity, and

speech content (among other types of information, see Campanella & Belin, 2007; Yovel & Belin, 2013). However, our results also hint at two potential areas of improvement in previous face-voice processing models (Belin et al., 2004; Campanella & Belin, 2007). First, while these models focus on three parallel pathways for both faces and voices devoted to emotion, speech, and identity recognition, they have not retained a fourth processing route from Bruce and Young's (1986) face processing model (on which these newer voice processing models were based): This particular "directed visual processing" route was dedicated to process age, gaze, and gender information in faces. Our present results suggest that similar, interdependent (e.g., "feature-selective processing") routes for (integrated) face and voice processing might be useful to capture the present face-voice gender congruency effects. Additionally, our results suggest that interactive processing in these pathways does not appear to require (top-down, or instruction-based) directed attention (Bruce & Young, 1986), but rather proceeds in a relatively automatic manner. Furthermore, cross-modal interaction in such "feature-selective processing" pathways appears to be able to directly affect the auditory speech processing (i.e., gender-incongruency between pathways made speech content harder to process), thereby emphasizing the interactive nature between different streams of face-voice processing (see Figure 3).

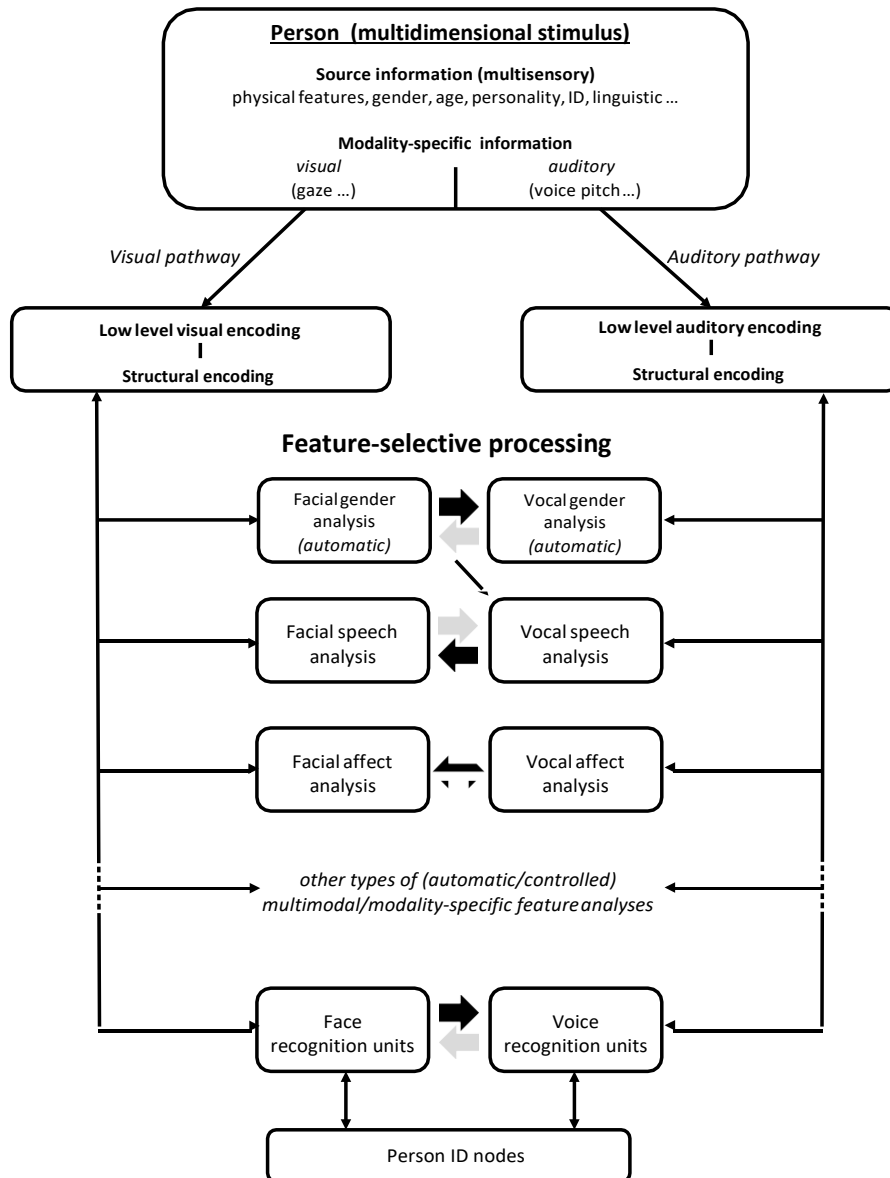


Figure 3. Schematic outline of face-voice processing based on previous models (see Bruce & Young, 1986; Belin et al., 2004 for general details). Here, any multi-/uni-modal feature or source information in the context of person perception can be subject of selective processing, including gender and speech processing. The model also highlights some well-known asymmetries in cross-modal interaction strength (grey/black arrows, e.g., see Huestegge & Raettig, 2018). The present study suggests that (relatively) automatic cross-modal interaction of gender processing affects vocal speech analysis (digit categorization).

Our somewhat surprising observation of a reversed congruency effect in RTs for switch trials in the task switching blocks might point to effects of processing prioritization in terms of capacity scheduling. Specifically, the presence of such an effect in the first place indicates that there was still sufficient cognitive capacity to compute gender incongruency information. However, given the extraordinarily difficult executive control demands in switch trials, participants appear to be aware of the additional difficulty associated with gender-incongruent trials, and thus particularly assign more capacity to the handling of these specific trials, ultimately resulting in relatively better performance (compared with gender-congruent conditions). This assumption seems especially plausible since it is mirrored by the observation that the magnitude task, which per se is clearly easier than the parity task as indicated by RTs and error rates pure block performance, is associated with longer RTs and higher error rates (as well as greater cross-task response congruency effects and greater switch costs, see Meuter & Allport, 1999, for similar effects in the context of switching to the easier of two languages) in the task switching blocks, also indicating that participants assign more capacity to more difficult conditions. Of course, this can only be a preliminary assumption that needs to be tested more extensively in future research.

At first sight, the attenuation of congruency effects under strong cognitive (executive control) load in task switching (vs. pure) blocks appears to be in conflict with some previous studies (e.g., Zäske et al., 2016, who implemented a quite different methodology), including a parallel study from our own lab which is based on the same stimulus material (Huestegge & Raettig, 2018). However, note that in the present study (unlike in Huestegge & Raettig, 2018) gender processing was *not* task-relevant, which might explain why an attenuation in the task switching (vs. pure) blocks was observed here.

Of course, it is difficult to assess the extent to which the present results (effects of task-irrelevant face-voice gender congruency on speech-based information processing) can be

generalized to information processing in general (in terms of effects of congruency between two task-irrelevant stimulus dimensions on processing of any third, task-relevant stimulus dimension). For example, the fact that the same source (the voice) carried both task-relevant and (at least one part of the) task-irrelevant information may have contributed to the observed congruency effects. Furthermore, it has been argued that faces and voices may be special because they are social stimuli that may capture attention (Theeuwes & Van der Stigchel, 2006) and tap into domain-specific attentional resources (Neumann & Schweinberger, 2009; Zäske et al., 2016). Therefore, future research is needed to determine the extent to which the present results may generalize to information processing in general.

The congruency effects that emerged in the present study suggest that we never listen to the speech content of voices in isolation, but always attend to the person and its gender as a whole (embodied processing). This may also have implications for the treatment of people whose voices are inauthentic in the sense that they are not sufficiently congruent with their visual appearance (including transgender people). Specifically, our data suggest that further support to increase the matching of visual and voice-related gender information (e.g., using voice training) may help to improve communication.

### References

- Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance series. Attention and performance 15: Conscious and nonconscious information processing* (pp. 421-452). Cambridge, MA: MIT Press.
- Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Frontiers Psychology, 15(5)*, 727.
- Barsalou L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-45.
- Barsalou, L. W., Simmons, W. K., Barbey, A., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*, 84-91.
- Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8(3)*, 129–135. Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11(12)*, 535–543.
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77(3)*, 305-327.
- Campeanu, S., Craik, F. I. M. & Alain, C. (2013). Voice Congruency Facilitates Word Recognition. *PLoS One. 8(3)*.
- Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88(1)*, 143-156.
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon identification of a target letter in a non- search task. *Perception and Psychophysics, 16*, 143-149.

- Freeman, J. B., & Ambady, N. (2011). When two become one: Temporally dynamic integration of the face and voice. *Journal of Experimental Social Psychology, 47*, 259-277.
- Geiselman, R. E., & Crawley, J. M. (1983). Incidental Processing of Speaker Characteristics - Voice As Connotative Information. *Journal of Verbal Learning and Verbal Behavior, 22*(1), 15-23.
- Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition 22*(5), 1166-83.
- Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50*(6), 524-536.
- Greeno, J. G., & Middle School Mathematics through Applications Project Group (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*(1), 5-26.
- Huestegge, S., & Raettig, T. (in press). Crossing gender borders: Bidirectional dynamic interaction between face-based and voice-based gender categorization. *Journal of voice*. <https://doi.org/10.1016/j.jvoice.2018.09.020>
- Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin, 136*(5), 849-874.
- Kim, S. Y., Kim, M. S., & Chun, M. M. (2005). Concurrent working memory load can reduce distraction. *Proceedings of the National Academy of Science of the United States of America, 102*, 16524-9.



- Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility – A model and taxonomy. *Psychological Review*, *97*, 253-270.
- Latinus, M., VanRullen, R., & Taylor, M. J. (2010). Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC Neuroscience* *11* (11), 36.
- Masuda, S., Tsujii, T., & Watanabe S. (2005). An interference effect of voice presentation on face gender discrimination task: Evidence from event-related potentials. *International Congress Series*, *1278*, 156-159.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *22*, 1423-1442.
- Meuter, R. F. I., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language*, *40*(1), 25-40.
- Murphy, G., Groeger, J. A., & Greene, C. M. (2016). Twenty years of load theory-Where are we now, and where should we go next? *Psychonomic Bulletin & Review*, *23*(5), 1316-1340.
- Neumann, M. F., & Schweinberger, S. R. (2009). N250r ERP repetition effects from distractor faces when attending to another face under load: Evidence for a face attention resource. *Brain Research*, *1270*, 64-77.
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355-376.
- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech-Perception As A Talker-Contingent Process. *Psychological Science*, *5*(1), 42-46.

- Palmeri T. J., Goldinger S. D., & Pisoni D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning Memory and Cognition*, *19*(2), 309-28.
- Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899-1910.
- Peynircioglu, Z., Brent, W., Tatz, J., & Wyatt, J. (2017). McGurk effect in gender identification: Vision trumps audition in voice judgments. *The Journal of General Psychology* *144*(1), 59-68.
- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207-231.
- Schweinberger, S. R., & Robertson, D. M. C. (2017). Audiovisual integration in familiar person recognition. *Visual Cognition*, *25*(4-6), 589-610.
- Schweinberger, S. R., Robertson, D. M. C., & Kaufmann, R. M. (2007). Hearing facial identities. *Quarterly Journal of Experimental Psychology*, *60*(10), 1446-1456.
- Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology*, *17*(19), 1680-1685.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643-662.
- Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657-665.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review* *9*(4), 625-636.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices.

*Trends in Cognitive Science*, 17(6), 263-271.

Zäske, R., Fritz, C., & Schweinberger, S. R. (2013). Spatial inattention abolishes voice

adaptation. *Attention Perception & Psychophysics*, 75(3), 603-613.

Zäske, R., Perlich, M. C., & Schweinberger, S. R. (2016). To hear or not to hear: Voice

processing under visual load. *Attention Perception & Psychophysics*, 78(5), 1488-

1495.

Zäske, R., Volberg, G., Kovacs, G., & Schweinberger, S. R. (2014). Electrophysiological

Correlates of Voice Learning and Recognition. *Journal of Neuroscience*, 34(33),

10821-10831.

### **Acknowledgements**

The authors would like to thank Vera Volk for her help in data acquisition, and Stephan Suschke and his Würzburg team of actors at Mainfranken Theater Würzburg for providing stimulus models.

# Study 1

Huestegge, S., & Raettig, T. Crossing gender borders: Bidirectional dynamic interaction between face-based and voice-based gender categorization

In press: *Journal of Voice* (accepted for publication on 25.09.2018)  
Published "online first": <https://doi.org/10.1016/j.jvoice.2018.09.020>

Bei Studie 1 handelt es sich um eine Version (post-review, pre-proof) des o.g. Artikels (Nutzung hier mit freundlicher Genehmigung des Elsevier-Verlags).

# Study 2

Huestegge, S. M., Raettig, T., & Huestegge, L. (2019). Are face-incongruent voices harder to process? Effects of face-voice gender incongruency on basic cognitive information processing. *Experimental Psychology*, 66, 154-164. (accepted for publication on 10.01.2019)

Bei Studie 2 handelt es sich um eine Version (post-review, pre-proof) des o.g. Artikels (Nutzung hier mit freundlicher Genehmigung von *Experimental Psychology* 2019; Vol. 66(2), 154–164 ©2019 Hogrefe Publishing; [www.hogrefe.com](http://www.hogrefe.com)).

DOI: <https://doi.org/10.1027/1618-3169/a000440>

# Study 3

Sujata M. Huestegge (subm.). Matching unfamiliar voices to static and dynamic faces: No evidence for a dynamic face advantage in a simultaneous presentation paradigm

Submitted: *Frontiers: Psychology* (15.02.2019)

Bei Studie 3 handelt es sich hier um eine Version (initially submitted version, but without any changes that occurred during the revision process) des o. g. Artikels.

Die finale Version wurde am 08.08.2019 von *Frontiers: Psychology* akzeptiert und kann unter DOI: [10.3389/fpsyg.2019.01957](https://doi.org/10.3389/fpsyg.2019.01957) (Copyright © 2019 Huestegge) abgerufen werden.

**Matching unfamiliar voices to static and dynamic faces:**

**No evidence for a dynamic face advantage in a simultaneous presentation paradigm**

Sujata M. Huestegge<sup>1,2</sup>

<sup>1</sup>University of Würzburg &

<sup>2</sup>University of Music and Performing Arts Munich

Correspondence address:

Sujata M. Huestegge

University of Music and Performing Arts Munich

Arcisstr. 12

80333 Munich, Germany

[sujata.huestegge@hmtm.de](mailto:sujata.huestegge@hmtm.de)

## Abstract

Previous research has demonstrated that humans are able to match unfamiliar voices to corresponding faces and vice versa. It has been suggested that this matching ability might be based on common underlying source information that has characteristic effects on both faces and voices. Some researchers have additionally assumed that dynamic facial information might be especially relevant to successfully match faces to voices. In the present study, we compared static and dynamic face-voice matching ability in a simultaneous presentation paradigm. Additionally, we implemented a procedure (matching additionally supported by incidental association learning) which allows us to reliably exclude participants that did not pay sufficient attention to the task. A comparison of performance between static and dynamic face-voice matching suggested a lack of substantial differences in matching ability for static and dynamic faces, suggesting that dynamic (as opposed to mere static) facial information does not contribute substantially to face-voice matching performance. Implications regarding the underlying mechanisms of face-voice matching are discussed.

Current theories of face and voice processing assume strong interactive processing between corresponding visual and auditory input modalities (e.g., Campanella & Belin, 2007). This claim is especially plausible given that similar types of information about a person are processed based on faces and voices, for example, age, gender, ethnicity, masculinity/femininity, health, speech content, personality, emotion etc. A famous example for the combined usage of both streams of information is the McGurk effect, showing that processing different speech input in both channels (e.g., seeing a face uttering /ga/ while hearing a voice uttering /ba/) can result in an illusory fused percept (e.g., of hearing /da/, see McGurk & MacDonald, 1976). However, under more natural conditions cross-modal processing interactions do not lead us on the wrong track. Instead, cross-modal information redundancies can, in the absence of experimenters dubbing “wrong” audio tracks to videos, be used to enhance processing of person-related information. For example, it has been shown to be easier to classify a face when it is accompanied with its voice (Joassin, Maurage, & Campanella, 2011). Additionally, a recent cross-modal priming study suggests that face and voice information is already integrated early in the processing stream to enhance recognition (Bülthoff & Newell, 2017). A reverse conclusion from this is that faces and voices share common source identity information, thus principally allowing for bidirectional inferences between voices and faces. For example, in a situation where a hotel employee calls our name at the airport to pick us up, we can use his/her voice features to come up with educated guesses as to which one of the faces surrounding us might belong to this voice. Indeed, previous research has demonstrated that the accuracy of matching novel faces to voices is substantially above chance level, thus corroborating the claim that faces and voices share common source identity information.



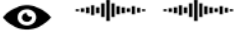



Lachs (1999) invented a sequential crossmodal matching task which was since then frequently used to study matching of static and dynamic faces to audio samples of voices (e.g., Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004 a,b; Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007). Such a crossmodal matching task typically consists of four sequential phases per trial: All stimuli (a reference stimulus in one modality prior to two sequentially presented comparison stimuli in the other modality) are presented sequentially, followed by a final decision phase (typically a two-alternative forced choice). The sequence of modalities can be varied (visual or auditory reference stimulus first), and the stimuli consisted either of single words (e.g., Lachs, 1999; Lachs & Pisoni, 2004) or sentences (e.g., Smith et al., 2016a; Krauss, Freyberg, & Morsella, 2002). Usually, such a design yielded only chance performance when static pictures of faces were used (but see Mavica & Barenholtz, 2013, Experiment 2, for a notable exception), but above-chance performance when dynamic faces were used as stimuli. Based on these findings, it has been argued that temporally dynamic facial information might be a necessary prerequisite for the ability to match faces and voices (e.g., Kamachi et al., 2003).

In the following, an alternative explanation for the discrepancy of results between static and dynamic faces has been proposed. It has been argued that the four-phase paradigm might impose a strong memory load, which could particularly affect the (potentially more demanding) static face condition (Smith, Dunn, Baguley, & Stacey, 2016a). Indeed, previous memory literature suggested that memory performance for dynamic faces is much better than for static faces (Christie & Bruce, 1998; Knappmeyer, Thornton, & Bülthoff, 2003; Lander & Chuang, 2005), probably due to the presence of more person-related cues (O'Toole, Roark, & Abdi, 2002). Furthermore, studies using simultaneous presentation of at least the comparison stimuli revealed clear above-chance matching performance (Krauss et al., 2002; Mavica & Barenholtz, 2013; Smith et al., 2016, Exp. 3). Among these studies, Krauss et al. (2002)

presented whole bodies of persons (instead of faces), and only Mavica and Barenholtz (2013, Experiment 1) used a completely simultaneous setting, in which the voice stimulus and two visual face stimuli were presented at once.

Still another matching paradigm involves the sequential presentation of two face-voice pairs, after which participants are asked to directly indicate in which of the two pairs the face matched the voice (same-different procedure, Smith, Dunn, Bagulay, & Stacey, 2016b). This paradigm also revealed above-chance matching performance. Following up on this, Smith, Dunn, Baguley, & Stacey (2018) confirmed that any delay between the presentation of the voice and the static face yields matching performance at chance level only. Recently, Stevenage, Hamlin, and Ford (2017) used a simultaneous same/different matching task and showed above-chance performance for matching voices to static faces. Additionally, they showed that the distinctiveness of the speaker's voice increased matching performance. Table 1 presents an overview of these previous methodologies.

Table 1. Overview of previous experimental procedures in face-voice matching studies.

	<b>Static</b>		<b>Dynamic</b>	
	<i>Matching one of two auditory stimuli to a visual stimulus</i>	<i>Matching one of two visual stimuli to an auditory stimulus</i>	<i>Matching one of two auditory stimuli to a visual stimulus</i>	<i>Matching one of two visual stimuli to an auditory stimulus</i>
				
<b>Sequential stimulus presentation</b>	V → A <sub>1</sub> → A <sub>2</sub> → R (Kamachi et al, 2003; Lachs & Pisoni, 2004 a; Lander et al 2007; Smith et al., 2016a)	A → V <sub>1</sub> → V <sub>2</sub> → R (Kamachi et al, 2003; Lachs & Pisoni, 2004 a; Lander et al 2007; Mavica & Barenholtz, 2013; Smith et al., 2016a)	V → A <sub>1</sub> → A <sub>2</sub> → R (Lachs 1999; Kamachi et al, 2003; Lachs & Pisoni, 2004 a,b; Lander et al 2007; Smith et al., 2016a)	A → V <sub>1</sub> → V <sub>2</sub> → R (Lachs 1999; Kamachi et al, 2003; Lachs & Pisoni, 2004 a,b; Lander et al 2007; Smith et al., 2016a)
<b>Semi-Simultaneous stimulus presentation</b>	Not feasible	A <sub>1</sub> → V <sub>1</sub> V <sub>2</sub> (Krauss et al., 2002; Smith et al., 2016a); A <sub>1</sub> V <sub>1</sub> → A <sub>1</sub> V <sub>2</sub> (Smith et al., 2016a)	Not feasible	A <sub>1</sub> V <sub>1</sub> → A <sub>1</sub> V <sub>2</sub> (Smith et al., 2016a)
<b>Simultaneous stimulus presentation</b>	Not feasible	A V <sub>1</sub> V <sub>2</sub> (Mavica & Barenholtz, 2013) → <b>Implemented in present study</b>	Not feasible	Possible, but not yet implemented → <b>Implemented in present study</b>

Note. V = Visual Stimulus (Face); A = Auditory Stimulus (Voice); R = Response in the absence of stimuli

In the present study, we designed an experiment involving both static and dynamic faces under optimal (i.e., simultaneous) and comparable conditions. Crucially, both conditions were implemented in a completely simultaneous paradigm in which a voice is accompanied with the display of either two static faces (as in Mavica & Barenholtz, 2013, Exp. 1), or with two dynamic faces, the latter representing a condition not reported until now, even though such a condition more realistically captures matching demands in daily life situations, such as those referred to in the beginning. Participants indicated via a left/right key press which of the two (left/right) faces matches the heard voice while all relevant information is still present (thereby preventing any delay between stimulation and response that was still present in many previous studies). Of course, it is not feasible to implement a modality-reversed condition

(matching one face to two simultaneously presented audio files) in such a simultaneous matching paradigm.

Taken together, there is common agreement that dynamic face information can be matched to voices in typical research paradigms that were previously used. Furthermore, there is also some empirical support for the ability to match static faces to voices, but the research so far might be taken as an indication that dynamic face-voice matching abilities are superior to static face-voice matching abilities. In the present study, we aim to replicate previous findings showing the ability to match dynamic faces to voices, but by using a novel, simultaneous presentation paradigm similar to the one used in previous research for static faces (Mavica & Barenholtz, 2013, Experiment 1). In the same paradigm, we test the ability to match both static and dynamic faces to voices in order to add further empirical evidence to the conflicting results in previous research, and to compare static and dynamic face-voice matching performance. A careful interpretation of the previous research methodologies suggests that simultaneous presentation should be especially advantageous for finding above-chance matching performance.

Importantly, we additionally reasoned that one potentially crucial issue in face-voice matching experiments can be the participant's motivation to comply with the task. That is, some participants may not be sufficiently motivated to actually try to match faces and voices, resulting in matching performance that does not differ significantly from chance level. These participants may sometimes (but not always) be characterized – and therefore identified – by particularly fast RTs or schematic response behavior (e.g., always pressing the same key or two keys in constant alternation). However, it nevertheless appears difficult to finally decide whether a participant had chance-level performance simply because he/she had weak face-voice matching abilities, or whether he/she did not comply with the instructions (or sufficiently attend to the task) in the first place (and should therefore be discarded from the

analysis). Unfortunately, in the crucial comparison between static and dynamic face matching performance this can be a quite serious issue, since an unequal number of such non-complying participants between conditions can yield serious performance difference artifacts between groups. We thus utilized an incidental learning design to be better able to control for participants that do not sufficiently pay attention to the task.

In the present specific design, another factor was therefore introduced (beside face-voice matching ability) that is also suited to support matching performance based on incidental association learning. Specifically, participants repeatedly encountered the same stimuli across two-choice matching trials with one correct matching option in each trial. This should eventually support an attentive participant's performance, irrespective of his/her actual face-voice matching abilities. Consider, for example, a Trial 1 in which Voice A is presented alongside Face X and Face Y (i.e., either X or Y must belong to A). If in a later trial (e.g., Trial 5) Voice A is presented again, but with Face Y and Face Z, participants could principally conclude that Face Y must belong to Voice A (since one option must be right, but Z cannot be a correct option since it was not presented in Trial 1). Overall, this should support matching performance, even though only to a limited extent, as participants are clearly not able to memorize all previous combinations and draw corresponding conclusions. Since the presence of face-voice matching abilities is already well established in the literature, we reasoned that it is not a severe problem that this procedure does not allow us to dissect the final performance into an incidental learning portion and face-voice matching ability portion. Instead, this procedure should allow us to judge whether participants pay attention to the task at hand and generally try to comply with instructions, assuming that any attentive participant should benefit from the incidental learning cues to achieve above-chance performance. Thus, we are able to exclude all participants that do not show above chance performance (here: equivalent to less than 100 correct trials out of 172, see method section for details), assuming

that these participants did not sufficiently pay attention to the task demands. Eventually, this should help us to only include attentive participants in both groups and thereby contribute to a better comparability between groups.

### Method

*Participants.* We originally aimed at testing 48 participants in the experiment (24 in the dynamic faces group and 24 in the static faces group). Participants were randomly assigned to either of the two conditions (static vs. dynamic faces). All participants had normal (self-reported) hearing and normal or corrected-to-normal vision. They received monetary reimbursement (or a small present) for participation. All participants gave informed consent and were treated in accordance with the Declaration of Helsinki (the study is considered exempt from a full ethic vote procedure from the local ethics committee).

In the dynamic group, 5 participants of 24 showed performance (in terms of errors) that was not significantly different from chance level (note that 4 of them were participants with the shortest mean RTs within their group, further corroborating the assumption that these participants did not pay close attention to the task). Thus, we filled up the sample with 5 new participants (performance of one participant did again not differ significantly from chance). In the static group, performance of one of the 24 participants did not differ significantly from chance (we decided against replacing this single data set with a new participant). Thus, each group finally consisted of 23 participants with above-chance performance (i.e., indicating attention to task, see above). In sum, the sample in the static group had a mean age of 25 years ( $SD = 4.5$ , range: 20-36 years, 4 left-handed, 7 male), while the dynamic group had a mean age of 24 years ( $SD = 4.2$ , range: 19-33 years, 2 left-handed, 6 male).

*Stimuli & Apparatus.* Participants were seated about 45 cm in front of a TFT computer screen (24") with a standard computer keyboard in front of them. Two keyboard keys (left Ctrl, Alt) served as response keys to indicate the visually presented face that matches the

voice (displayed on the left/right of the screen). The stimuli were based on original video and audio recordings of 8 male and 6 female speakers (professional theater actors). Since we aimed at natural visual face presentation, the videos also showed the hair and upper torso of the actors. Fourteen middle-aged actors (age range: 29-53 years, dialect-free native speakers, neutral facial expression, unambiguous male/female face and voice) were selected as stimulus models. We deliberately avoided to include young and elderly persons as stimuli to minimize the potential of age serving as a strong cue for matching performance. Stimuli were based on an audio file with a mean duration of 22.1 s ( $SD = 2.9$ ) across all speakers, and represented a standardized short German text (sample from “Der Nordwind und die Sonne”, approx. translation: “the north wind and the sun”). At the same time, participants either saw two pictures of faces on the left and right side of the screen (static face condition), or two (silent) video clips of faces on the left and right side of the screen (neutrally) uttering a proverb (“aus einer Mücke einen Elefanten machen”, approx. equivalent to “make a mountain out of a molehill”, dynamic face condition). The two faces on the screen were always of the same gender. In the dynamic face condition, the proverb was looped, thereby the corresponding moving faces were visible throughout the trial. The mean duration of a single utterance (single loop) amounted to 1.9 s ( $SD = 0.25$ ). The correlation of the models’ speech rate between the (visually presented) proverb and the (auditorily presented) text amounted to  $r = .70$ ,  $p = .006$ , indicating that fast proverbs speakers also tended to be fast text speakers. The looped videos were edited using the software Pinnacle (Version 21.0 Ultimate, CorelDRAW®, Ottawa, Ontario, Canada), while the experiment was programmed using the software Presentation (Version 19.0, Neurobehavioral Systems, Inc., Berkeley, CA). The image/video combinations were displayed (presented side-by-side adjacent to the left and right of the screen center, each picture/video subtending  $30^\circ$  horizontally and  $17^\circ$  vertically) on black background.

*Procedure.*

The experiment started with an instruction screen. In each trial, after a brief presentation of a central fixation cross, participants were asked to attend to the heard voice and the two (static or dynamic) faces and to press the (left/right) key corresponding to the respective face that, according to the participant’s suggestion, best matched the voice. The stimuli were presented as long as participants gave their response; speed was not emphasized in the instruction. After each response, presentation of the stimuli finished and the next trial started (see Figure 1). Participants in each group completed 172 trials. Eight male speakers, combined with all respective other male models, resulted in 56 trial combinations. Each combination was presented twice (switching left and right positions), resulting in 112 male trials altogether. The same was done with the female models, resulting in  $30 \times 2 = 60$  female trials. Male and female trial combinations were presented in random order. No performance feedback was given. After completing the matching task, participants filled out a qualitative post-experiment enquiry, involving questions regarding the basis on which they derived their choices.

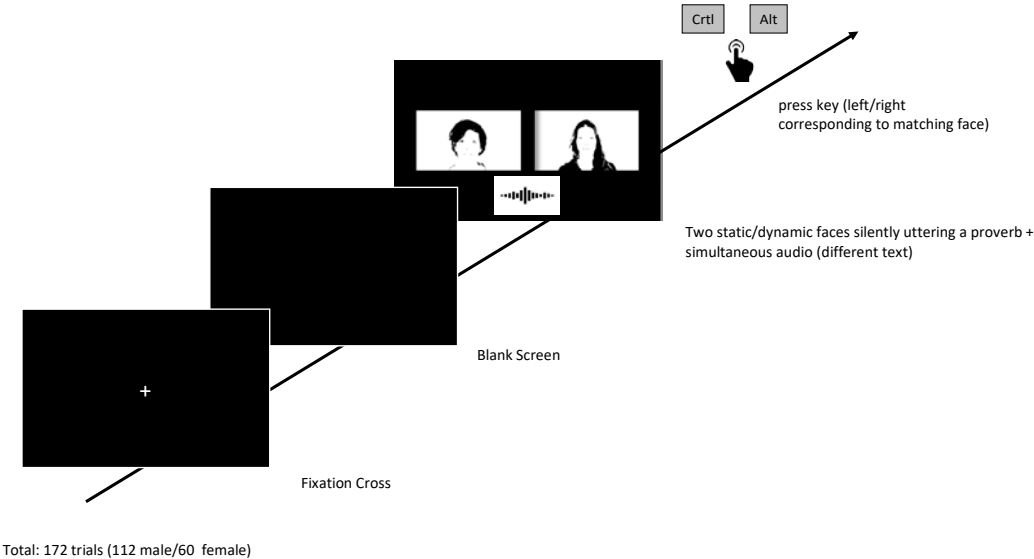


Figure 1. Schematic trial sequence.



## Results

*Matching performance.* The error rate amounted to 23.51% ( $SD = 7.52$ ) in the static group and 24.27% ( $SD = 7.53$ ) in the dynamic group,  $t(44) = 0.342$ ,  $p = .734$ . The 95% *CI* for the mean group difference of 0.76% ranged from -3.7% to 5.2%. Even at the extreme *CI* border (-3.7%) this difference clearly speaks against the assumption that additional dynamic (in contrast to mere static) information substantially contributed to matching performance.

*Response Times.* RTs < 600 ms (equivalent to two trials overall) as well as RTs > 6000 ms (equivalent to 8.3% of trials) were removed as outliers from RT analyses (note that this removal did not change any of the following result patterns). Mean RTs were 2465 ms ( $SD = 576$ ) in the static group and 2617 ms ( $SD = 530$ ) in the dynamic group,  $t(44) = 0.935$ ,  $p = .355$ . In the static group, error rates negatively correlated with RTs ( $r = -.668$ ,  $p < .001$ ), whereas this was not the case in the dynamic group ( $r = .248$ ,  $p = .254$ ). The difference between these two correlations was statistically significant,  $p < .05$ .

*Item-based analyses.* We also analyzed individual voice stimuli regarding their associated matching performance. There was no significant between-group RT difference for any of the voice stimuli, all  $ps > .05$ . Regarding error rates, two female stimuli (out of all 14 stimuli) significantly differed in error rates between groups (both  $ps$  between .01 and .05), but these two differences pointed into opposite directions. Overall, performance for individual stimuli was therefore quite comparable between the two groups.

There were significant matching performance differences (evidenced by error rates) between the male stimuli, both in the static,  $F(7, 154) = 9.187$ ,  $p < .001$ , and in the dynamic,  $F(7, 154) = 5.244$ ,  $p < .001$ , condition. This was also found for the female stimuli,  $F(5, 110) = 2.940$ ,  $p = .028$ ,  $F(5, 110) = 5.684$ ,  $p = .001$ , respectively. These effects indicate that some voices were easier to match with faces than others.

Error rates for the voice stimuli were significantly correlated across groups,  $r = .667$ ,  $p = .009$ , suggesting that voice stimuli that were (relatively) easy to match with static faces were also easy to match with dynamic faces (individual error rates for male voice stimuli in the static group: 29%, 20%, 13%, 9%, 23%, 9%, 26%, 32%, female: 27%, 36%, 31%, 19%, 26%, 38%, in the dynamic group: 23%, 17%, 19%, 9%, 29%, 16%, 18%, 28%, female: 17%, 31%, 31%, 39%, 31%, 49%, respectively). A qualitative analysis of these performance indices in the context of the age of the stimulus models (male: 29, 46, 34, 46, 41, 53, 35, 36 years, female: 35, 37, 32, 28, 47, 34 years) did not reveal any clear evidence that stimulus models at the edges of their respective age distribution were particularly easy to match (which would indicate that age served as a central cue for matching abilities): While the oldest male stimulus model was indeed associated with best performance in the static group, this was no longer the case in the dynamic group. In turn, the youngest male stimulus model was actually associated with the *worst* performance in the static group (and still among the worst in the dynamic group). For the female stimulus models, there was also no clear indication for age as a central cue for matching performance: The oldest female stimulus model was neither associated with best performance in the static nor in the dynamic group. While the youngest stimulus model was indeed associated with best performance in the static condition, she was associated with relatively low performance in the dynamic condition (where the best performance was associated with a stimulus model whose age was exactly the mean age of the female stimuli). Taken together, there was no consistent indication that age was a highly informative cue for matching abilities (note, however, that the number of stimuli was too low to conduct meaningful quantitative, statistical analyses).

*Initial matching performance (without incidental learning).* Even though the present design is not suited to reliably assess matching performance without the additional impact of incidental learning of matching patterns (see introduction), we additionally analyzed only the

first trial of each participant to assess matching accuracy without any contribution of incidental association learning. When combining both groups, a one-sided binomial test revealed statistically significant above-chance matching performance ( $N = 46$ , observed proportion: 65% correct, 35% incorrect,  $p = .027$ ). Of course, corresponding group-wise analyses suffer from severe power limitations. Nevertheless, these group-wise analyses still revealed above-chance performance at least in the dynamic group ( $N = 23$ , 74% correct,  $p = .017$ ), but not in the static group ( $N = 23$ , 57% correct,  $p > .05$ ). Overall, this analysis further confirms the central assumption that participants were able to match faces and voices, thereby replicating previous studies (see introduction).

*Qualitative questionnaire analysis.* We finally qualitatively analyzed the participants' verbal answers to the question regarding their subjectively experienced criteria to complete the matching decisions. Specifically, we were interested in the participants' awareness regarding the incidental learning support inherent in the present design. However, only 2 participants in the static group and 3 participants in the dynamic group mentioned that they somehow made use of the repeated presentation of faces and voices (in different combinations) as a cue to improve performance.

## Discussion

Taken together, the present study – which utilized a novel procedure to ensure attention to task based on implicit learning support – finds no evidence for any meaningful incremental usability of dynamic information in face-voice matching, as evidenced by the narrow borders of the corresponding confidence interval (95% CI) regarding the performance difference between static and dynamic face conditions. This result challenges earlier assumptions postulating that dynamic facial cues are of particular importance for matching faces to voices: Specifically, it has been argued that temporally dynamic facial information might be a necessary prerequisite for the ability to match faces and voices (e.g., Kamachi et

al., 2003). The present study is also the first to compare static and dynamic faces in an ideal situation for face-voice matching by simultaneously presenting all stimuli at the same time to reduce any memory-related additional demands (for sequential paradigms see: Lachs, 1999; Kamachi et al., 2003; Lachs & Pisoni, 2004 a,b; Lander et al., 2007). Note that a methodological strength of the present incidental learning approach is that we found a way to control for participants lacking sufficient attention to task demands, a phenomenon that has the potential for yielding strong artifacts in comparisons of static and dynamic face-to-voice matching performance (i.e., whenever the occurrence of such participants slightly differs between groups and therefore creates spurious group differences). Interestingly, this procedure initially led to more exclusions of participants in the dynamic condition, not in the static condition. Thus, whatever the information used by participants to complete the task (e.g., hormone levels expressed in both face and voice etc.), it should mainly be already present in the static face, and only to a negligible extent in dynamic (speaking-related) information.

Previous studies that have not revealed above-chance static face-voice matching ability were likely hampered by the strong memory load associated with sequential stimulus presentation paradigms, which particularly could affect the (potentially more demanding) static face condition (Smith et al., 2016a). This reasoning is also in line with previous memory literature which suggests that memory performance for dynamic faces is much better than for static faces (Christie & Bruce, 1998; Knappmeyer et al., 2003; Lander & Chuang, 2005). Notably, studies that used a simultaneous presentation of at least the comparison stimuli revealed clear above-chance matching performance for static face stimuli (Krauss et al, 2002; Mavica & Barenholtz, 2013; Smith et al., 2016a, Exp. 3), which is in line with the observations in the present study.

The present results are also important as they replicate previous reports of successful face-voice matching ability with both dynamic and static faces, but with a different set of stimuli. Many of the previous reports only rely on a few sets of selected stimuli, and it is therefore important to replicate corresponding effects using new sets of faces and voices.

However, the present approach also has several potential limitations which should be discussed. First, the incidental learning procedure does not provide a separate estimate of face-voice matching ability (i.e., unconfounded with a general ability to implicitly learn cross-modal stimulus associations). At first sight, one might therefore challenge the assumption that face-voice matching ability substantially contributes to performance in the present task in the first place. However, this argument is not compatible with the analysis of only the first trial in each participant, which (when taking both groups into account) clearly revealed significant matching abilities in the present sample of participants. Also note that there was no indication of a ceiling effect in the performance of the participants, and there was a reasonable extent of between-participants variability. Again, these observations further support the claim that incidental learning did not override any face-voice-matching-related effects.

Another potential issue is that there was no statistically significant evidence for static face-voice matching ability when analyzing only the first trials in these 23 participants (as opposed to a significant corresponding effect in the dynamic condition). At first sight, this observation seems to corroborate suggestions from previous research that matching ability is enhanced for dynamic stimuli. However, several observations speak against this conclusion. First, full performance (across all trials) was highly comparable between both groups, and there is no reason to believe that in the static face group incidental association learning ability was much stronger and therefore somehow compensated for any reduced face-voice matching ability (note that participants were randomly assigned to groups). Taken together, the most likely explanation for the lack of a significant face-voice matching effect in the static group

using the first trial per participant only is simply a lack of statistical power for this particular analysis. Finally, the assumption of significant static face-voice matching ability is also confirmed by previous studies demonstrating corresponding effects (Krauss et al., 2002; Mavica & Barenholtz, Experiment 2; Smith et al., 2016, Exp. 3), and some studies suggested that the absence of strong memory demands is important when assessing static face-voice matching ability (Smith et al., 2016b; Smith et al., 2018; Stevenage et al., 2017).

Finally, another potential issue in the present design is that the visual presence of two talking faces (uttering a proverb) in addition to a voice that utters an unrelated text may have yielded some level of confusion in participants from the dynamic faces group that might be lower in the static faces group. Of course, it is principally possible that this has prevented participants from exhibiting better performance in the dynamic condition. However, any measure to *not* present dynamic (and static, for the sake of comparability) faces and voices simultaneously would yield other, potentially more crucial disadvantages for the comparison of static and dynamic face conditions (see corresponding arguments presented earlier).

Nevertheless, further research should demonstrate static face-voice matching ability with the present stimulus set without using an incidental association learning support, for example, by using fewer trials (without stimulus repetitions) and a larger number of participants to increase statistical power. The present results (as well as previous studies, see above) strongly suggest that significant matching performance should be observed under these conditions. If possible, it would also be advisable to use an even narrower age range for the stimulus models to further minimize any possibility of using perceived age as a cue to increase matching performance, even though there was no clear evidence in the data that stimulus age was used as a cue for matching faces to voices.

Our results are overall in line with current theories of face and voice processing that assume a strong interaction between corresponding visual and auditory input modalities (e.g.,

Campanella & Belin, 2007). The assumption of a strong interaction is especially plausible given that similar types of information about a person are processed based on faces and voices. However, given the overwhelming evidence for matching ability in the literature, more future research should be devoted to address the underlying mechanisms of this ability in terms of the specific sources of the correlates between voice and face features that participants use to match faces and voices (e.g., hormone levels affecting both faces and voices in a predictable manner). Such “concordant information” includes a variety of potential characteristics (see Smith et al., 2016a, for a review). For example, faces and voices both contain information regarding genetic fitness and attractiveness, masculinity/femininity (based on hormone levels), age, health, height, and weight (see also Yehia, Rubin, & Vatikiotis-Bateson, 1998). The present results suggest that some relevant sources should already affect static face characteristics, and are not predominantly expressed in facial movements while speaking.

Interestingly, Nagrani, Albanie, and Zisserman (2018) recently presented a computer algorithm that was claimed to even exceed face-voice matching abilities of humans, merely based on physical features of (static and dynamic) visual face information and voice audio (not taken from the same video material). Again, this corroborates the common source information hypothesis and opens up interesting technical possibilities of exploring such common features across modalities in technical applications in the near future.

## References

- Bülthoff, I., & Newell, F. N. (2017). Crossmodal priming of unfamiliar faces supports early interactions between voices and faces in person perception. *Visual Cognition, 25*, 611-628.
- Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*, 535-543.
- Christie, F., & Bruce, V. (1998). The role of dynamic information in the recognition of unfamiliar faces. *Memory & Cognition, 26*, 780-790.
- Joassin, F., Maurage, P., & Campanella, S. (2011). The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study. *NeuroImage, 54*, 1654-1661.
- Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology, 13*, 1709-1714.
- Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43*, 1921-1936.
- Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618-625.
- Lachs, L. (1999). A voice is a face is a voice: Cross-modal source identification of indexical information in speech. In *Research on spoken language processing* (Progress Report No. 23, pp. 241-258). Bloomington, IN: Indiana University, Department of Psychology, Speech Research Laboratory.
- Lachs, L., & Pisoni, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology, 16*, 159-187.



- Lachs, L., & Pisoni, D. B. (2004b). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*, *116*, 507-518.
- Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*, 429-442.
- Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: Matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 905-914.
- Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 307-312.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- Nagrani, A., Albanie, S., & Zisserman, A. (2018). Seeing Voices and Hearing Faces: Cross-modal biometric matching. *Processings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- O'Toole, A. J., Roark, D., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science*, *6*, 261-266.
- Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016a). Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, *78*, 868-879.
- Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016b). Concordant cues in faces and voices: Testing the back-up signal hypothesis. *Evolutionary Psychology*, *14*, 1474704916630317.

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2018). The effect of inserting an inter-stimulus interval in face–voice matching tasks. *The Quarterly Journal of Experimental Psychology*, *71*, 424-434.

Stevenage, S., Hamlin, I., & Ford, B. (2017). Distinctiveness helps when matching static faces and voices. *Journal of Cognitive Psychology*, *29*, 289-304.

Yehia, H. C., Rubin, P. E., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*, 23-44.

#### Acknowledgements

I would like to thank Tim Raettig for his support in programming the experiment, and Vera Volk and Miriam Kottmann for their help in testing the participants. I finally thank Lynn Huestegge for scientific writing support and helpful discussions about the design, analyses, and data interpretation.

# Anhang

(Material)

Einverständniserklärung (Probanden)

Nachbefragungsbogen der Studie 1

Nachbefragungsbogen der Studie 2

Nachbefragungsbogen der Studie 3

Stimuli der gesamten Studien auf CD-Rom:

288 synchronisierte audiovisuelle Videosequenzen (Zahlwort; ca. 1 s Dauer)

14 stumme Videos (Sprichwort; 25-30 s-loop)

14 Bilddateien (aus obigem Videomaterial extrahiert)

14 Audio-Dateien (Text; ca. 20-25 s)

## Einverständniserklärung zur Studienteilnahme (Huestegge\_Raettig\_2018)

Sehr geehrte Probandin, sehr geehrter Proband,

vielen Dank für Ihre Teilnahme an dieser Studie.

Alle in dieser Studie gesammelten Daten werden anonymisiert und nach DSGVO-Richtlinien erhoben und ausgewertet und ausschließlich von Fachleuten für wissenschaftliche Zwecke verwendet. Die Ergebnisse der Studie werden anonymisiert in einer Gruppe zusammengefasst und ohne Bezug auf konkrete Personen wissenschaftlich veröffentlicht.

Ihre Teilnahme an dieser Studie ist freiwillig. Sie haben jederzeit die Möglichkeit, die Studie ohne Angabe von Gründen abubrechen. Entsprechend Ihrer investierten Zeit erhalten Sie eine anteilige Entlohnung in Form von VP-Stunden. Die Einwilligung zur Verwendung Ihrer Daten können Sie während der Studienteilnahme jederzeit widerrufen. Ein nachträglicher Widerruf nach Beendigung der Studie ist aufgrund der anonymisierten Speicherung Ihrer Daten nicht möglich.

Angaben zur Person (Entsprechendes bitte einkreisen / ergänzen):

Geschlecht:  
männlich / weiblich

Alter:  
\_\_\_\_\_

Händigkeit:  
rechts / links

Sichtigkeit:  
\_\_\_\_\_

- Ich habe die aufgeführten Bedingungen gelesen und verstanden. Ich bin über Ablauf und Zweck der Studie unterrichtet und eventuelle Fragen sind durch den Versuchsleiter ausreichend beantwortet worden. Ich hatte genügend Zeit, eine Entscheidung zu treffen. Mit meiner Unterschrift bestätige ich mein Einverständnis zur Teilnahme an dieser Studie.

Name (in Druckschrift):

\_\_\_\_\_

Datum:

\_\_\_\_\_

Unterschrift:

\_\_\_\_\_



VP-Nr.: \_\_\_\_\_ **Nachbefragung Experiment 1; 2; 1 und 2 (switch)**

**Kennen Sie eine der dargebotenen Personen? ja \_\_\_\_\_ nein \_\_\_\_\_**

**1.1 Was glauben Sie: Worum ging es im Experimentteil „Gesichter erkennen“?**

**1.2 Wie schwer fanden Sie das Experiment (1 = sehr leicht, 7 = sehr schwer)?**

○-----○-----○-----○-----○-----○-----○  
1        2        3        4        5        6        7

**1.3 Welche Strategie haben Sie zur Bewältigung der Aufgabe verfolgt?**

**2.1 Was glauben Sie: Worum ging es im Experimentteil „Stimme erkennen“?**

**2.2 Wie schwer fanden Sie das Experiment (1 = sehr leicht, 7 = sehr schwer)?**

○-----○-----○-----○-----○-----○-----○  
1        2        3        4        5        6        7

**2.3 Welche Strategie haben Sie zur Bewältigung der Aufgabe verfolgt?**

**3.1 Was glauben Sie: Worum ging es im Experimentteil „Wechselaufgabe“?**

**3.2 Wie schwer fanden Sie das Experiment (1 = sehr leicht, 7 = sehr schwer)?**

○-----○-----○-----○-----○-----○-----○  
1        2        3        4        5        6        7

**2.3 Welche Strategie haben Sie zur Bewältigung der Aufgabe verfolgt?**

**3. Haben sie bei jeder Aufgabe darauf achten können, immer die Gesichter anzuschauen? Ja  Nein**



VP-Nr.: \_\_\_\_\_ **Nachbefragung Experiment 3; 4; 3 und 4 (switch)**

**Kennen Sie eine der dargebotenen Personen? ja \_\_\_\_\_ nein \_\_\_\_\_**

**1.1 Was glauben Sie: Worum ging es im Experimentteil „größer/kleiner“?**

**1.2 Wie schwer fanden Sie das Experiment (1 = sehr leicht, 7 = sehr schwer)?**

○-----○-----○-----○-----○-----○-----○  
1        2        3        4        5        6        7

**1.3 Welche Strategie haben Sie zur Bewältigung der Aufgabe verfolgt?**

**2.1 Was glauben Sie: Worum ging es im Experimentteil „gerade/ungerade“?**

**2.2 Wie schwer fanden Sie das Experiment (1 = sehr leicht, 7 = sehr schwer)?**

○-----○-----○-----○-----○-----○-----○  
1        2        3        4        5        6        7

**2.3 Welche Strategie haben Sie zur Bewältigung der Aufgabe verfolgt?**

**3.1 Was glauben Sie: worum ging es im Experimentteil „Wechselaufgabe“?**

**3.2 Wie schwer fanden Sie das Experiment (1 = sehr leicht, 7 = sehr schwer)?**

○-----○-----○-----○-----○-----○-----○  
1        2        3        4        5        6        7

**2.3 Welche Strategie haben Sie zur Bewältigung der Aufgabe verfolgt?**

**3. Haben sie bei jeder Aufgabe darauf achten können, immer die Gesichter anzuschauen?** Ja  Nein



