Received: 11 February 2021 | Revised: 26 April 2021 | Accepted: 2 May 2021

DOI: 10.1002/hbm.25532

RESEARCH ARTICLE

Visual mechanisms for voice-identity recognition flexibly adjust to auditory noise level

Corrina Maguinness[1,2] | Katharina von Kriegstein[1,2]

[1]Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, Technische Universität Dresden, Dresden, Germany

[2]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Correspondence
Corrina Maguinness, Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, Technische Universität Dresden, Dresden, Germany.
Email: corrina.maguinness@tu-dresden.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: KR 3735/5-1; H2020 European Research Council, Grant/Award Number: SENSOCOM 647051; Max Planck Research Group Grant, Grant/Award Number: awarded to KvK

Abstract

Recognising the identity of voices is a key ingredient of communication. Visual mechanisms support this ability: recognition is better for voices previously learned with their corresponding face (compared to a control condition). This so-called 'face-benefit' is supported by the fusiform face area (FFA), a region sensitive to facial form and identity. Behavioural findings indicate that the face-benefit increases in noisy listening conditions. The neural mechanisms for this increase are unknown. Here, using functional magnetic resonance imaging, we examined responses in face-sensitive regions while participants recognised the identity of auditory-only speakers (previously learned by face) in high (SNR -4 dB) and low (SNR $+4$ dB) levels of auditory noise. We observed a face-benefit in both noise levels, for most participants (16 of 21). In high-noise, the recognition of face-learned speakers engaged the right posterior superior temporal sulcus motion-sensitive face area (pSTS-mFA), a region implicated in the processing of dynamic facial cues. The face-benefit in high-noise also correlated positively with increased functional connectivity between this region and voice-sensitive regions in the temporal lobe in the group of 16 participants with a behavioural face-benefit. In low-noise, the face-benefit was robustly associated with increased responses in the FFA and to a lesser extent the right pSTS-mFA. The findings highlight the remarkably adaptive nature of the visual network supporting voice-identity recognition in auditory-only listening conditions.

KEYWORDS
audio-visual, FFA, motion, multisensory, predictive coding, pSTS, voice-identity

1 | INTRODUCTION

Human communication is often based on input from more than one sensory modality. For example, when we listen to someone's voice, we are often concurrently exposed to their face. These audio-visual correspondences make communication more robust. For instance, in noisy listening conditions, observers can more accurately perceive what someone says when they can also view their accompanying lip-movements (Erber, 1969; Rosenblum, Johnson, & Saldana, 1996; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954). Perhaps surprisingly, these visual mechanisms are also engaged under auditory-only listening conditions (Sheffert & Olson, 2004; von Kriegstein et al., 2008; von Kriegstein & Giraud, 2006): Listeners are more accurate at recognising the identity of a speaker by their voice alone, when that speaker has been previously learned by face, in comparison to a control condition (Schall, Kiebel, Maess, & von Kriegstein, 2013; Schelinski, Riedel, & von Kriegstein, 2014; Sheffert & Olson, 2004; von Kriegstein et al., 2008; Zäske, Mühl, &

Schweinberger, 2015). This behavioural enhancement, termed the 'face-benefit', emerges rapidly following approximately 2 min of audio-visual experience with the speaker's identity (von Kriegstein et al., 2008). The face-benefit is observed in the majority of neurotypical participants (i.e., 76%—von Kriegstein et al., 2008; Maguinness, Schall, & von Kriegstein, 2021) and might be one of the reasons why most of us recognise familiar voices with such ease (Lavan, Burton, Scott, & McGettigan, 2019; Maguinness, Roswandowitz, & von Kriegstein, 2018; Maguinness & von Kriegstein, 2017; Sidtis & Kreiman, 2012; Stevenage, 2018).

Audio-visual learning likely benefits unisensory processing as the information in each sensory stream is governed by a *common* cause (for reviews see Shams & Seitz, 2008; von Kriegstein, 2012). Voices are caused by physical visual structures (i.e., the vocal tract) and provide information about the visual characteristics of the speaker. For example, fundamental frequency (i.e., pitch), formant frequencies and vocal-tract resonance (i.e., timbre) map well to, and are predictive of structural form cues, including face-identity (Ghazanfar et al., 2007; Ives, Smith, & Patterson, 2005; Kim et al., 2019; Krauss, Freyberg, & Morsella, 2002; Mavica & Barenholtz, 2013; Oh et al., 2019; Smith, Dunn, Baguley, & Stacey, 2016a; Smith, Dunn, Baguley, & Stacey, 2016b; Smith & Patterson, 2005; Smith, Patterson, Turner, Kawahara, & Irino, 2005). This non-arbitrary coupling of sensory information is reflected at the neural level: The face-benefit for auditory-only voice-identity recognition has been shown to be mediated by responses in the fusiform face area (FFA; Schall et al., 2013; von Kriegstein et al., 2008; von Kriegstein & Giraud, 2006). The FFA is a visual face-sensitive region (Kanwisher, McDermott, & Chun, 1997) implicated in the processing of structural facial form (i.e., the invariant static features of the face) and face-identity (Axelrod & Yovel, 2015; Eger, Schyns, & Kleinschmidt, 2004; Ewbank & Andrews, 2008; Grill-Spector, Knouf, & Kanwisher, 2004; Kanwisher & Yovel, 2006; Liu, Harris, & Kanwisher, 2010; Rotshtein, Henson, Treves, Driver, & Dolan, 2005; Schiltz, Dricot, Goebel, & Rossion, 2010; Weibert & Andrews, 2015; Xu, Yue, Lescroart, Biederman, & Kim, 2009). Responses in this region, during voice-identity compared to speech recognition, occur as early as 110 ms after auditory onset (Schall et al., 2013)—a time point when voice-identity recognition has yet to be achieved (Schweinberger, 2001; Schweinberger, Kloth, & Robertson, 2011). This quick response is thought to be mediated by direct connections between the FFA and voice-sensitive regions in the superior temporal gyrus and sulcus (STG/S) (Blank, Anwander, & von Kriegstein, 2011; Hölig, Föcker, Best, Röder, & Büchel, 2014a, 2014b; Schall & von Kriegstein, 2014; von Kriegstein & Giraud, 2006; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). Similar early audio-visual processing mechanisms also occur when the face and voice are presented concurrently, or in succession (Föcker, Hölig, Best, & Röder, 2011; Schweinberger et al., 2011).

What might be a governing principle for cross-modal interactions during auditory-only tasks? One proposal offered by an audio-visual model of human auditory communication (review see von Kriegstein, 2012; von Kriegstein et al., 2008, 2005; von Kriegstein & Giraud, 2006) is that visual mechanisms assist auditory recognition by generating predictions about, and thus placing constraints on the sensory processing of, the incoming auditory signal (Blank, Kiebel, & von Kriegstein, 2015; Kiebel, Daunizeau, & Friston, 2009; von Kriegstein et al., 2008). Such a process would be particularly beneficial for optimising voice-identity recognition when the auditory signal is weak or degraded, with predictions assisting recognition by 'filling in' missing sensory information. Although concerned with unisensory auditory processing, this proposal is reminiscent of the principle of inverse effectiveness (Meredith & Stein, 1986) when concurrent multimodal inputs are available, that is, enhanced multisensory integration when the saliency of the unimodal inputs is weak. In agreement with the model's proposal (von Kriegstein, 2012; von Kriegstein et al., 2008; von Kriegstein et al., 2005), recent behavioural evidence shows that, in individuals who display a face-benefit, the face-benefit for voice-identity recognition increases with decreasing signal-to-noise ratios (SNRs) of the auditory signal (Maguinness et al., 2021). This indicates that learned visual mechanisms may help to systematically resolve incoming noisy auditory input. While previous studies have shown a positive relationship between increased FFA responses and the face-benefit in relatively clear listening conditions (von Kriegstein et al., 2008), to-date, it is unclear whether the face-benefit for voice-identity processing in noise is also facilitated by responses in the FFA.

Voice-identity recognition is mediated by the extraction of relatively invariant 'static' voice features, such as fundamental frequency and vocal tract resonances (Latinus & Belin, 2011; Lavner, Rosenhouse, & Gath, 2001; Voiers, 1964). However, there is also evidence that *dynamic* articulatory idiosyncrasies, such as speech rhythm (Dellwo, Leemann, & Kolly, 2015; He & Dellwo, 2016; Leemann, Kolly, & Dellwo, 2014; Van Lancker, Kreiman, & Emmorey, 1985; Van Lancker, Kreiman, & Wickens, 1985) and formant dynamics (Ingram, Prandolini, & Ong, 1996; Mc Dougall, 2004, 2006; Mc Dougall & Nolan, 2007; Zuo & Mok, 2015) play a role. These dynamic cues can support voice-identity recognition when other cues such as fundamental frequency are unreliable (Fellowes, Remez, & Rubin, 1997; Remez, Fellowes, & Rubin, 1997; Sheffert, Pisoni, Fellowes, & Remez, 2002; Simmons, Dorsi, Dias, & Rosenblum, 2021; Zuo & Mok, 2015). In parallel, similar adaptive mechanisms have also been observed to support face-identity recognition when static form cues are degraded. In challenging viewing conditions, facial motion cues or 'dynamic facial signatures' (O'Toole, Roark, & Abdi, 2002; Roark, Barrett, Spence, Abdi, & O'Toole, 2003) can provide a complementary route to recognition (Dobs, Bülthoff, & Schultz, 2016; Knight & Johnston, 1997; Lander & Bruce, 2000; Lander, Christie, & Bruce, 1999; Lander & Chuang, 2005; Longmore & Tree, 2013). These cues are likely processed by motion-sensitive regions of the face-network (Bernstein & Yovel, 2015; Girges, O'Brien, & Spencer, 2016; Girges, Spencer, & O'Brien, 2015; O'Toole et al., 2002), that is, the posterior superior temporal sulcus motion-sensitive face area or pSTS-mFA (Bernstein, Erez, Blank, & Yovel, 2018; Fox, Iaria, & Barton, 2009; Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011; Schultz & Pilz, 2009). Like static cues, dynamic spatio-temporal cues in the face and voice share common source identity information (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs &

Pisoni, 2004; Mc Dougall, 2006; Smith et al., 2016a, 2016b; Simmons et al., 2021). Thus, in conditions with noise, the face-benefit for voice-identity recognition might rely on complementary dynamic face-identity cues processed in the pSTS-mFA, rather than the FFA. Such a finding would indicate that stored visual cues may be used in an adaptable manner, in line with the nature of the auditory input, to support voice-identity processing (Figure 1).

The aim of the present study was to investigate the visual mechanisms underpinning the face-benefit for voice-identity recognition in noisy listening conditions (Figure 1). We used functional magnetic resonance imaging (fMRI) to examine responses in both the FFA and the pSTS-mFA while participants engaged in an auditory-only voice-identity recognition task in two levels of auditory noise: high-noise (signal-to-noise ratio −4 dB) and low-noise (signal-to-noise ratio +4 dB). All speakers had been learned before MRI-data acquisition. Crucially, half of the speakers had been learned by seeing and listening to videos of the speaker talking (voice-face learning), while the other half had been learned by listening to the speaker while viewing a visual control stimulus depicting the speaker's occupation (voice-occupation learning). Thus, the design was a 2 × 2 factorial design with the factors noise-level (high-noise, low-noise) and learning condition (voice-face, voice-occupation). Our first and central aim was to test whether in the high-noise listening condition (in contrast to the lower noise condition) there are increased responses in the FFA, the pSTS-mFA, or both for face-learned (in comparison to occupation-learned) speakers, that is, a noise-level x learning condition interaction. Second, we expected a positive correlation between listeners' face-benefit scores and responses in visual face-sensitive regions, that is, a behaviourally relevant relationship with neural responses. Third, we expected that face-sensitive regions underpinning the face-benefit in noise would share functional connections with voice-sensitive regions in the STG/S (Figure 1). On the behavioural level, we expected based on previous findings (Sheffert & Olson, 2004; for review see von Kriegstein, 2012), that across both noise levels, speakers learned by face would be more accurately recognised than those learned by occupation and that this face-benefit would be greatest in the high-noise condition (Maguinness et al., 2021).

2 | MATERIALS AND METHODS

2.1 | Participants

Twenty-three neurotypical German speaking adults (12 female; mean age 25 years, *SD* 2.9 years), recruited from the Max Planck Institute for Human Cognitive and Brain Sciences participant database, took part in this study. We did not employ a formal power analysis prior to the start of the study. The current sample size is similar to previous

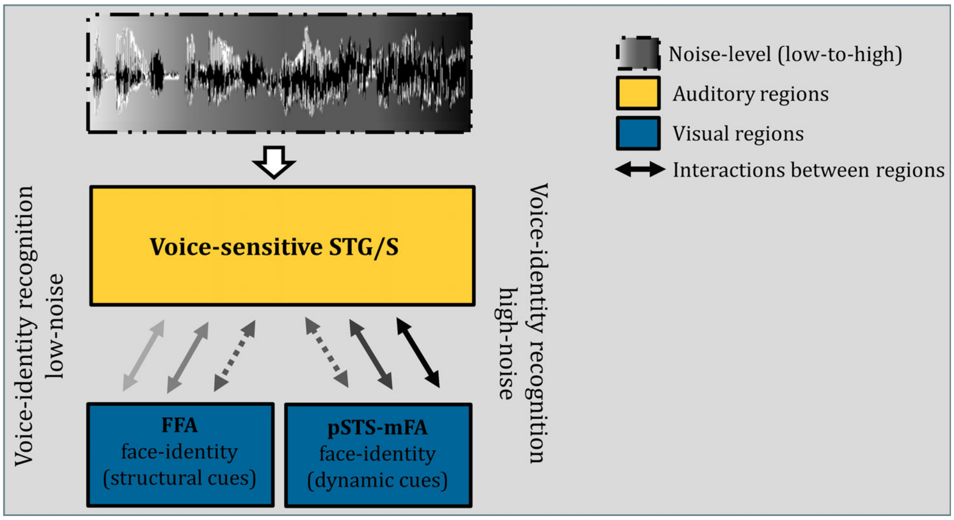

FIGURE 1 Schematic overview of potential audio-visual interactions between voice- (yellow) and face-sensitive (blue) brain regions, during auditory-only voice-identity processing. The interactions may vary as a function of the noise level present in the auditory signal (top panel low-to-high noise, light grey-to-black transitioning colours). Interactions between regions are indicated via bidirectional arrows. The arrow colours (light grey-to-black) reflect the noise level (low-to-high) in the auditory signal. Bold arrows depict potential strong interactions between brain regions, dashed arrows depict weaker interactions. Voice-identity recognition low-noise. Voice-identity recognition is supported by interactions between the FFA, a region sensitive to structural face-identity cues, and the anterior and mid voice-sensitive STG/S (Schall & von Kriegstein, 2014; von Kriegstein & Giraud, 2006). To date, this has been documented for relatively low-noise listening conditions for example, MRI-scanner noise, or in conditions with a positive signal-to-noise ratio (bold light grey/grey arrows, left of figure). The FFA may also support voice-identity processing in increasingly noisy listening conditions, however, the region may be less recruited as static vocal cues become increasingly degraded (dashed grey arrow, left of figure). Voice-identity recognition high-noise. The pSTS-mFA may be involved in voice-identity recognition in increasingly noisy listening conditions. Potentially, interactions between the pSTS-mFA, a region sensitive to dynamic face cues, and voice-sensitive regions in the anterior and mid STG/S, may be observed. This may be particularly apparent in high-noise levels (bold dark grey/black arrows, right of figure), and less so in lower noise levels (dashed grey arrow, right of figure)

studies investigating the face-benefit and associated neural responses (Blank et al., 2015; Schall et al., 2013; Schelinski et al., 2014; von Kriegstein et al., 2008; von Kriegstein & Giraud, 2006). All were right handed (Oldfield, 1971) and reported normal hearing and normal vision. All participants gave their informed written consent prior to participation according to the procedures approved by the Ethics Committee of the Medical Faculty at the University of Leipzig (299–12-24092012). Two subjects were excluded from neuroimaging and behavioural analysis owing to below chance performance on the voice-identity recognition task inside the MRI-machine. Analysis for the fMRI and behavioural task data was based on 21 participants (12 female; mean age 25 years, *SD* 3 years). Since one additional participant did not complete the face area localiser runs (see below), analysis of the fMRI functional face area localiser data was based on 20 participants.

2.2 | Stimuli

2.2.1 | Stimuli for the audio-visual training

The stimuli for the audio-visual *voice-face* and *voice-occupation* training sessions comprised of 10 audio-visual and five auditory-only recordings of 14 five- to six-word sentences, from six male German speakers (22–27 years old). All sentences were semantically neutral (e.g., 'Die Ente kommen an das Ufer' English: 'The ducks come to the shore'). Audio-visual stimuli for the voice-face training were video sequences, which displayed the talking face of the speaker. For the audio-visual voice-occupation training the speaker's face was replaced with an image depicting the speaker's occupation. Both audio-visual sequences were presented for the same duration.

The stimuli were recorded using a high-definition camera (Legria HF S10 HD-Camcorder, Canon, Japan) and an external condenser microphone [TLM 50 (Neumann, Berlin, Germany); Mic-Preamp, Mic-Amp F-35 (Lake People, Konstanz, Germany); soundcard, Power Mac G5 (Apple Inc., CA); Sound Studio 3 (Felt Tip, Inc. NY) (44.1 kHz sampling rate and 16 bit resolution)]. For the audio-visual training, video stimuli were edited in Final Cut Pro software (Apple Inc., CA) to include a circular mask, which excluded the background while revealing the face of the speaker. Videos were cropped to 727 × 545 pixels. In addition, for each speaker a single frame depicting the speaker in a neutral pose was extracted from the video sequence. Three symbols representing an occupation (painter, chef, and mechanic) were taken from Clip Art (http://office.microsoft.com/en-us/). The auditory stimuli were adjusted for overall mean amplitude using Matlab7 (MathWorks, MA).

2.2.2 | Stimuli for the auditory-only voice-identity recognition test

The stimuli for the auditory-only voice-identity recognition test (in the MRI-machine) consisted of 30 two-word sentences, presented in noise. Each sentence started with 'Er' (English: 'He') and finished with a verb (e.g., 'Er beisst', English: 'He bites'). All sentences were spoken by the same six male speakers presented during the audio-visual training phase.

The stimuli were recorded using the same apparatus as the stimuli for the audio-visual training. The stimuli were adjusted for overall mean amplitude using Matlab7 (MathWorks, MA) and then masked with pink noise (created in Matlab7 by filtering Gaussian white noise). Pink noise was chosen as it has similar spectral qualities to speech and has been used in previous studies examining the face-benefit on auditory processing (e.g., Maguinness et al., 2021; Schall et al., 2013) and audio-visual speech-in-noise processing (Riedel, Ragert, Schelinski, Kiebel, & von Kriegstein, 2015; Ross et al., 2007). Unlike white noise, it also has a stronger power in the frequency range (100–250 Hz) which is sensitive to spectral components of the speech signal that support identity-recognition that is, fundamental frequency (F0) (Pernet & Belin, 2012; Traunmüller & Eriksson, 1994). The stimuli were mixed with noise of varying intensities to produce signal-to-noise ratios (SNR) of -4 dB (referred to hereafter as 'high-noise') and $+4$ dB (referred to hereafter as 'low-noise'). The noise was ramped and introduced with a linear 50 ms fade-in and fade-out. For four participants included in the neuroimaging and behavioural analyses the auditory stimuli in the high-noise condition had an SNR of -8 dB and the low-noise $+4$ dB, these SNRs were chosen based on a previous behavioural study examining the face-benefit in noise, which used 4 dB interval steps ranging from SNR -8 dB to SNR $+4$ dB (Maguinness et al., 2021). Example auditory stimuli can be viewed in Figure 2. Sample sound files are available at: https://osf.io/d52c8/. The high-noise condition was subsequently adjusted to SNR -4 dB for the remaining participants in order to improve behavioural task performance inside the MRI-machine. The four participants could reliably complete the task (i.e., above chance performance). See Figure S1 (Supporting Information), for individual behavioural data. As we were interested in the within-subject effect of noise-level (i.e., relative effect of noise), all 21 participants were included in the analysis.

2.2.3 | Stimuli for the visual-only face area localiser

We used a functional localiser (Borowiak, Maguinness, & von Kriegstein, 2019; design as von Kriegstein et al., 2008) to establish the location of the face-sensitive FFA and the pSTS-mFA within participants. The face stimuli consisted of still frames, extracted using Final Cut Pro software (Apple Inc., CA), from video sequences of 50 identities (25 female; 19–34 years). All identities were unfamiliar and had no overlap with the identities from the main experiment. The video sequences were recorded using a digital video camera (HD-Camcorder LEGRIA HSF100; Canon Inc., Tokyo, Japan). In each sequence, the person was asked to stand still and look into the camera (frontal face view) with a neutral expression. In addition, each person articulated the letters of the German alphabet, maintaining the neutral pose. The object stimuli were static images of 50 different common

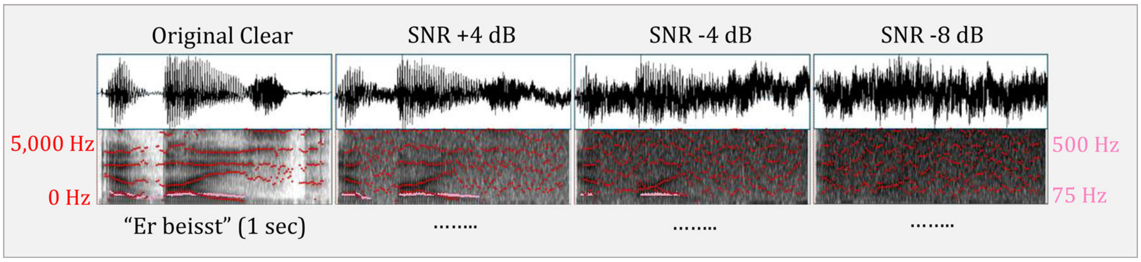

FIGURE 2 Example of the stimuli from the auditory-only voice-identity recognition task. Stimuli for one sample sentence, 'Er beisst' ('he bites'), are shown for signal-to-noise ratios of +4 dB, −4 dB, and −8 dB. The original clear audio file is shown on the left for comparison purposes. The spectrogram (lower panel) displays the fundamental frequency (F0), that is, pitch of the voice, in pink and the formant frequencies (F1–F5) in red

objects, which were taken from the database of object images described in (von Kriegstein et al., 2008). All images were presented in colour and cropped to measure 768 × 576 pixels. The sequence order and presentation rate of the multiple still frames were manipulated to ensure that the face and object images were perceived to be either static (i.e., stream of images of individual faces or objects) or dynamic (i.e., one person or one object moving onscreen) in nature (see Visual-only face area localiser (fMRI) for full details).

2.3 | Experimental procedure

All experiments including the audio-visual training, auditory-only voice-identity recognition test, and visual-only face area localiser were run using Presentation (www.neurobs.com) software.

2.3.1 | Audio-visual training (before MRI-data acquisition)

Prior to MRI data acquisition, participants were familiarised with six male speakers using an established audio-visual training paradigm which has been shown to elicit the face-benefit and associated responses in the FFA (see Schall et al., 2013; von Kriegstein et al., 2008). As we were specifically interested in the effect of noise level on these responses, we kept the learning design as comparable as possible to the previous studies—both of which have used male speaker sets. However, we note that previous studies examining voice-identity recognition of personally familiar voices (male and female voices), also demonstrate FFA responses (von Kriegstein et al., 2005; von Kriegstein, Kleinschmidt, & Giraud, 2006).

During the audio-visual training, three of the speakers were learned through an audio-visual sequence which displayed the corresponding dynamic facial identity of the speaker (i.e., video). The other three speakers were learned through an audio-visual control sequence, which displayed a visual image of the occupation of the speaker (Figure 3a). The inclusion of an audio-visual, rather than an auditory-only, control condition ensured that participants were always exposed to person-related visual information during learning. We refer to the two audio-visual training conditions as voice-face learning and voice-occupation learning, respectively. The three speakers assigned to the voice-face learning or the voice-occupation learning conditions were counterbalanced across participants. In both conditions, the participant also learned the name of the speaker.

Familiarisation with the speakers was achieved through two training rounds. Each training round consisted of a learning stage, followed by an evaluation stage. In the learning stage, participants were exposed to a series of trials which first displayed the name of the speaker (1 s), immediately followed by an audio-track (approx. 2 s). The audio-track in each trial consisted of one of 10 five- to six-word sentences (e.g., 'Die Enten kommen an das Ufer' English: 'The ducks come to the shore'). In the voice-face learning condition the audio-track was accompanied by the corresponding time synchronised video track. In the voice-occupation learning condition it was accompanied by the occupation symbol of the respective speaker. The voice-face or voice-occupation trials were presented in two separate blocks. There were 20 trials per speaker in each block that is, each speaker was heard uttering each five- to six-word sentence twice (i.e., 20 trials × 3 speakers), with a total of 60 trials per voice-face or voice-occupation block. In each block, the initial 15 trials were presented grouped by a speaker-identity that is, 5 trials were presented consecutively per speaker-identity. In the remaining 45 trials, all speaker-identities were presented in a randomised order. In total, there were 120 trials for a learning stage (60 voice-face trials, 60 voice-occupation trials). At the end of the learning stage, participants then completed an evaluation stage. In each evaluation trial, participants heard the voice of a speaker (auditory-only), which was immediately followed by a name (half of trials) or a static face/occupation image (half of trials) presented onscreen. The auditory clips consisted of four novel five- to six-word sentences which were not contained within the learning stage. In the whole evaluation stage, each speaker uttered each five- to six-word sentence twice, so that there were 8 evaluation trials per speaker (i.e., 24 trials for the voice-face and 24 trials for the voice-

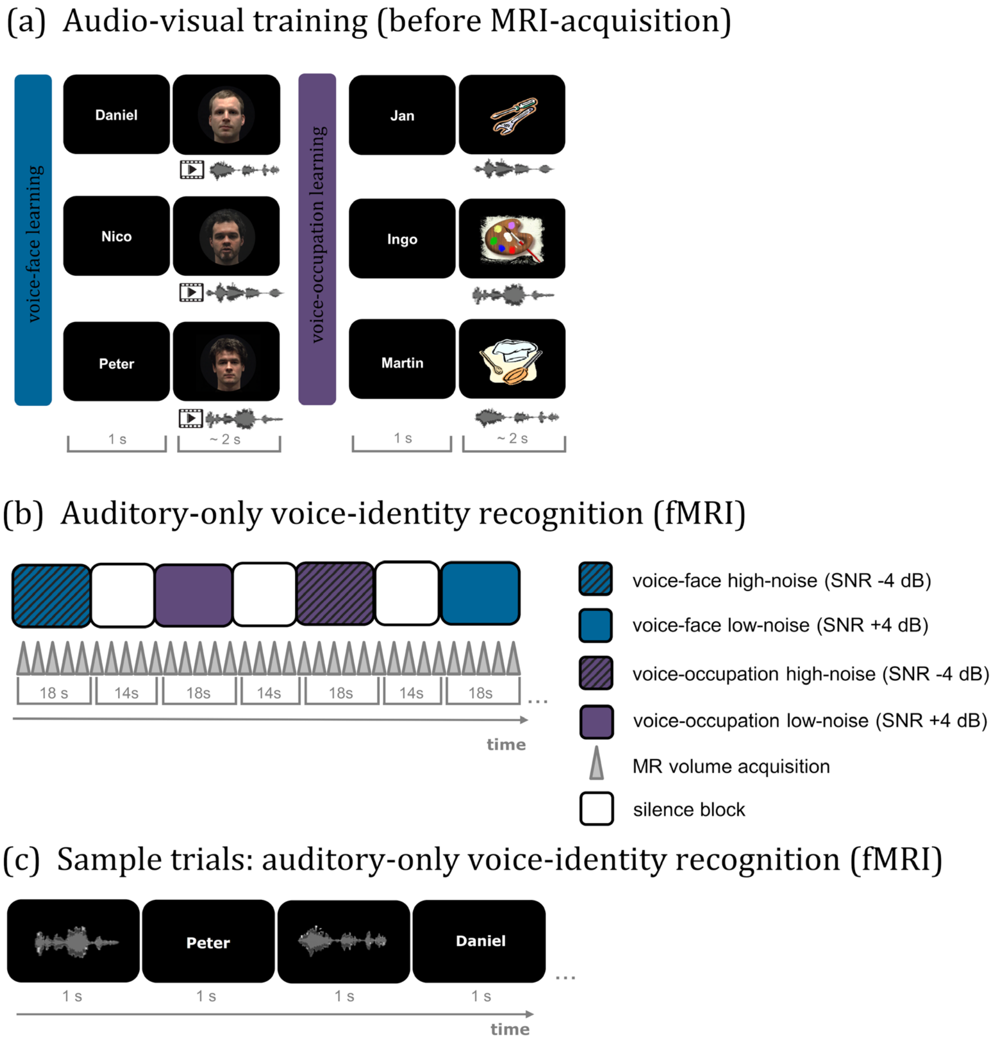
(a) Audio-visual training (before MRI-acquisition)
voice-face learning
Daniel
Nico
Peter
voice-occupation learning
Jan
Ingo
Martin
1 s
~ 2 s
1 s
~ 2 s
(b) Auditory-only voice-identity recognition (fMRI)
18 s
14s
18s
14s
18s
14s
18s
...
time
voice-face high-noise (SNR -4 dB)
voice-face low-noise (SNR +4 dB)
voice-occupation high-noise (SNR -4 dB)
voice-occupation low-noise (SNR +4 dB)
MR volume acquisition
silence block
(c) Sample trials: auditory-only voice-identity recognition (fMRI)
Peter
Daniel
1 s
1 s
1 s
1 s
...
time

FIGURE 3 A schematic illustration of the audio-visual training phase and the auditory-only voice-identity recognition test. (a) Audio-visual training. Prior to MRI-acquisition, participants learned the voice and name of six speakers. Half of the speakers were learned in conjunction with their corresponding face i.e., video (voice-face learning) and the other half with an occupation symbol (voice-occupation learning). The speakers assigned to each learning condition were counterbalanced across participants. Each speaker was learned for approximately 2 min in total.
(b) Auditory-only voice-identity recognition (fMRI). During MRI-acquisition, participants listened to auditory-only sentences spoken by the familiarised six speakers in high- and low-noise listening conditions. The speakers were presented in separate blocks, blocked by learning type (voice-face or voice-occupation) and noise level. Blocks were presented in a randomised order and interleaved with silence baseline blocks. Functional MR images were acquired continuously. (c) Sample trials: auditory-only voice-identity recognition (fMRI). On each trial, participants heard a speaker utter a sentence, followed by the presentation of a speaker's name onscreen. Participants decided whether the name matched the identity of the preceding voice. Note that the face-identities shown in (a) are for illustration purposes, and some differ from those used in the audio-visual training phase. These images are not displayed due to consent restrictions

occupation condition, totalling 48 trials for an evaluation stage). On half of the trials, for each learning condition, the voice-identity matched the name or static face/occupation image, while the other half of the trials contained mismatched names and images. The mismatching names and images were always taken from the same speaker set (i.e., voice-face or voice-occupation). Participants indicated via button-press whether the voice-identity matched the name or face/occupation image (‘yes’ or ‘no’). The name or image remained onscreen until a response was made. Participants received feedback in the form of the correct name and voice-face/voice-occupation combination. All participants reached the learning criterion of ≥80% correct (von Kriegstein et al., 2008) after two rounds of training, indicating that they could reliably match the correct combinations of voice, name, and face/occupation.

2.3.2 | Auditory-only voice-identity recognition test (fMRI)

In the auditory-only voice-identity recognition test (Figure 3b), participants listened to two-word sentences (e.g., ‘Er liest’, English: ‘He reads’). The sentences were uttered by the 6 speakers which participants had been familiarised with during the audio-visual training phase. In each trial, a sentence (1 s duration) was followed by a visual name presented onscreen (1 s duration; Figure 3c). When the visual name was displayed, the participant was instructed to respond 'yes' or 'no' as to whether the name matched the identity of the heard speaker. Participants indicated their response by pressing one of two assigned buttons with their right hand (responses were made within the 1 second name display interval). Half of the trials contained speakers who had been previously learned by face in the audio-visual training phase (voice-face learning), while the other half contained speakers who had been learned with an occupation symbol (voice-occupation). In addition, half of the trials were presented in high-noise and half in low-noise listening conditions. Thus, the experiment was a 2 × 2 factorial design with the factors learning (voice-face, voice-occupation) and noise-level (high-noise, low-noise). The experimental trials were blocked by condition type: (a) voice-face high-noise; (b) voice-face low-noise; (c) voice-occupation high-noise; (d) voice-occupation low-noise. Each block contained nine trials (18 s). There were 20 blocks per condition (720 trials in total). The blocks were presented in a randomised order and were interleaved with baseline silence blocks (14 s), in which participants looked at a fixation cross. The blocks were presented over four 11-min runs, with 20 task blocks per run.

2.3.3 | Visual-only face area localiser (fMRI)

We used a standard experiment to localise the FFA and the pSTS-mFA (Borowiak et al., 2019; Pitcher et al., 2011; von Kriegstein et al., 2008). Participants were presented with still frames taken from videos of faces or objects under four different conditions (Figure 4a): (a) images of faces from *different* identities, with different facial speech poses; (b) images of faces from the *same* identity, with different facial speech poses; (c) images of *different* objects, from different viewpoints; (d) images of the *same* object, from different viewpoints. All images were static in nature and presented onscreen for 500 ms with no interstimulus interval (Figure 4b). This fast image presentation rate induced an implied motion effect for images of the *same* facial identity (dynamic face) and for images of the *same* object (dynamic object) that is, the images appeared as one moving speaking face or one moving object onscreen (Figure 4b, dynamic conditions). This implied motion effect was not apparent when images depicted different facial identities (static face) or different objects (static object) (Figure 4b, static conditions). Images were blocked by condition type and each block contained 50 images. There were four blocks per condition type and each block lasted 25 s. The blocks were presented in a randomised order over two 6-min runs (8 blocks per run) and interleaved with baseline blocks, where a fixation cross was presented for 18 s

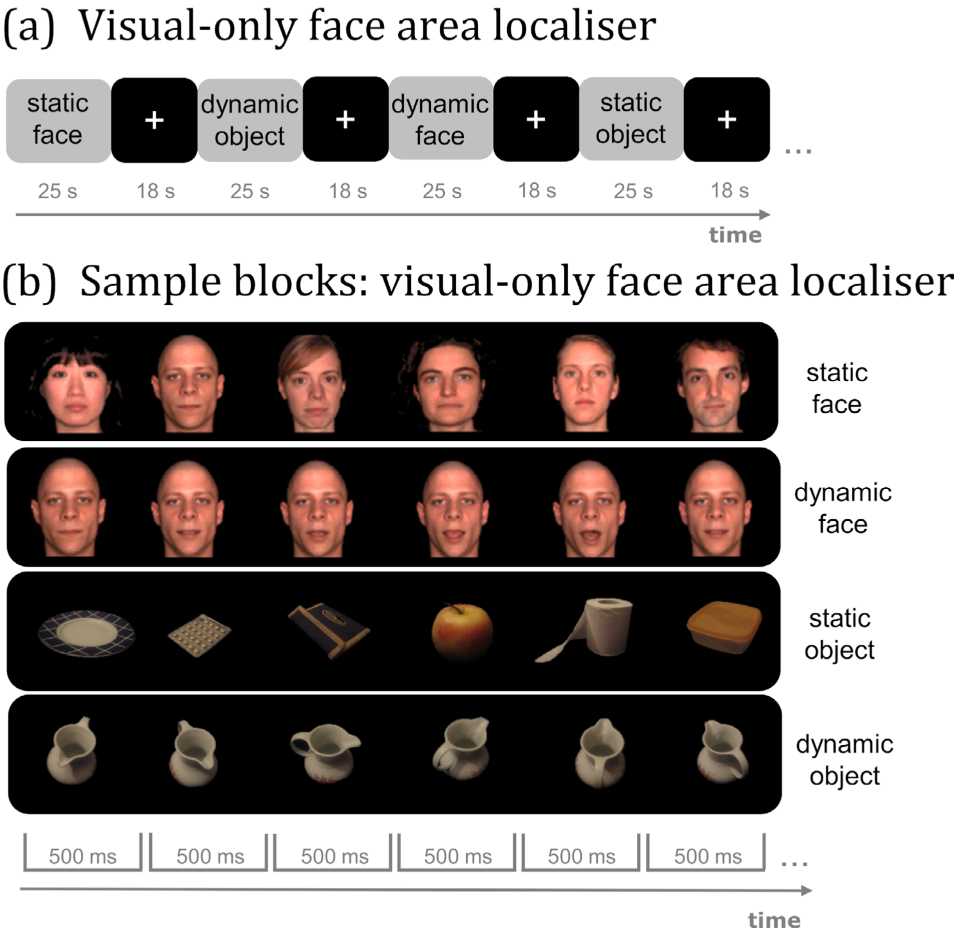

FIGURE 4 A schematic illustration of the (a) visual-only face area localiser. Images of faces and objects were shown in separate blocks, interleaved with baseline blocks displaying a fixation cross. There were four block types i.e., conditions: static face, dynamic face, static object, and dynamic object. Participants were asked to attend to the images presented onscreen. (b) Sample blocks: visual-only face area localiser. Sample stimuli and block structure for the four different conditions. Each image in a block was presented for 500 ms, with no interstimulus interval

(Figure 4a). Participants were asked to attentively view the images within each block.

2.4 | Image acquisition

2.4.1 | Functional MRI

Functional images for the auditory-only voice-identity recognition task and the visual face area localiser were acquired on a 3 T Siemens Prisma MR scanner (Siemens, Erlangen, Germany), equipped with a 20-channel head coil. Images were acquired using a gradient-echo echo planar imaging continuous scanning sequence (TE 30 ms, flip angle 90°, TR 2.64 s, 42 slices, whole-brain coverage, acquisition bandwidth 128 kHz [2004 Hz/pix], 2.5 mm slice thickness, 0.5 mm inter slice gap, in plane resolution 3 × 3 mm, ascending interleaved slice acquisition). Geometric distortions were characterised by a B0 field map. The field map scan consisted of a pair of 2D gradient echo images with different echo times (TE1/TE2 = 4.58 ms/7.04 ms). This field map was acquired once per participant before the first experimental task run began. All images were acquired in AC-PC orientation. Nine hundred eighty volumes were acquired for the auditory-only voice-identity recognition task (245 × 4 runs) and two hundred sixty-two volumes for the visual face area localiser (131 × 2 runs).

2.4.2 | Structural MRI

Structural images for each participant were attained from the MPI brain database. The images had been acquired on either the same 3 T Siemens Prisma scanner used for functional image acquisition, or a 3 T Siemens-Tim Trio, Magnetom Verio, or Numaris 4 scanner. Images were acquired using either a T1-weighted three-dimensional magnetization-prepared rapid gradient echo (MP-RAGE) sequence or a magnetization-prepared 2 rapid gradient echo (MP2-RAGE) sequence (MP-RAGE: $N = 13$; MP2-RAGE: $N = 8$). Imaging parameters for the MP-RAGE sequence were TR = 2,300 ms, TE = 2.98 ms, TI = 900 ms, flip angle = 9°, FOV = 256 mm × 240 mm, voxel size = 1 mm^3, 176 sagittal slices. Imaging parameters for the MP2-RAGE sequence were TR = 5,000 ms, TE = 2.92 ms, TI_1/TI_2 = 700 ms/2,500 ms, flip angle_1/flip angle_2 = 4°/5°, FOV = 256 mm × 240 mm, voxel size = 1 mm^3, 176 sagittal slices. All structural images were acquired using a 32-channel head coil, except for two participants where a 20-channel coil was used.

3 | DATA ANALYSIS

3.1 | Behavioural

Trials in which the participant failed to make a response, that is, missed trials, were disregarded from analysis (6% of trials). The overall trial count was then adjusted to include only trials on which a response was made. Accuracy was calculated as the number of correct responses divided by the adjusted trial count, for each participant, for each condition. Reaction times (in milliseconds) were also calculated for correct response trials, for each participant, for each condition. Behavioural data (voice-identity recognition performance: accuracy and reaction time) were analysed in Statistica (TIBCO Software) using a 2 × 2 repeated measures analysis of variance (ANOVA), with 'learning' (voice-face, voice-occupation) and 'noise-level' (high-noise, low-noise) as repeated factors. Effects were considered significant if present at $p < .05$. Effect sizes were calculated using partial eta square ${\eta_p}^2$ (Cohen, 1969; Richardson, 2011). A post-hoc power analysis conducted on the behavioural data ($N = 21$) demonstrated an achieved power of 0.6 and 0.9 for detecting a true effect of learning in the high- and low-noise conditions respectively (one-tailed, α error probability .05).

3.2 | Functional MRI

Functional MRI data were analysed with the statistical parametric mapping software package (SPM12, Wellcome Trust Centre for Neuroimaging, UCL, UK, (www.fil.ion.ucl.ac.uk/spm). We used standard spatial pre-processing procedures: images were realigned and unwarped, normalised to Montreal Neurological Institute (MNI) standard stereotactic space using the structural image of each participant, written to the original resolution 3 × 3 × 3 mm, and smoothed with an isotropic Gaussian filter of 8 mm at FWHM. Geometric distortion due to susceptibility gradients were corrected by an interpolation procedure based on the B0 field-map. Statistical parametric maps were generated by modelling the evoked hemodynamic response for the different conditions as boxcars convolved with a synthetic hemodynamic response function in the context of the general linear model (Friston, Ashburner, Kiebel, Nichols, & Penny et al., 2007). All contrasts of interest were computed at the single-subject level and then taken to a group-level random-effects analysis which estimated the second-level t-statistic at each voxel.

3.2.1 | Regions of interest

Visual regions of interest (ROI) for the functional response and connectivity analyses were the FFA (Kanwisher et al., 1997) and the pSTS-mFA (Pitcher et al., 2011). We localised the FFA with the contrast 'faces > objects'. The FFA was localised at the group level in the right hemisphere—maxima at $x = 45$, $y = -40$, $z = -19$ (T-value = 3.54). For the FFA ROI, the localiser was thresholded with a cluster size of 25 voxels. This cluster size is similar to previous reports for the FFA using this design and contrast (Borowiak et al., 2019; von Kriegstein et al., 2008). We did not observe responses in an analogous region in the left hemisphere at a threshold of <0.01 uncorrected. This right hemisphere dominance is in line with previous observations (Kanwisher et al., 1997; von Kriegstein et al., 2008). The pSTS-mFA was localised using the contrast 'dynamic faces > dynamic objects' (Fox et al., 2009). This contrast revealed a facial motion sensitive cluster in the right pSTS, with the maximum at the group level at $x = 54$, $y = -34$, $z = 2$ (T-value $= 4.55$) and an analogous cluster in the left hemisphere with the maximum at $x = -51$, $y = -46$, $z = 11$ (T-value $= 3.89$). Both regions were thresholded to have a comparable cluster size of 29 voxels. All three visual ROIs were also localised using data from a sub-sample of those 16 participants that showed a behavioural face-benefit to facilitate sub-group analyses. Identical contrasts, as described above, were used to identify these ROIs in this group ($N = 15$; one participant did not complete the localiser, see Section 2.1; Table S1 Supporting Information).

The auditory ROIs for the functional connectivity analysis were the voice-sensitive regions in the right middle and anterior STG/S (Belin & Zatorre, 2003; Schall, Kiebel, Maess, & von Kriegstein, 2014; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; von Kriegstein & Giraud, 2004). We chose these regions as they have been shown to share structural (Blank et al., 2011) and functional (Schall & von Kriegstein, 2014; von Kriegstein et al., 2005; von Kriegstein et al., 2006; von Kriegstein & Giraud, 2006) connections with the FFA. We defined the mid and anterior STG/S using spheres (8 mm radius) positioned around previously published co-ordinates in the right hemisphere for these regions (Blank et al., 2011): mid STG/S: $x = 63$, $y = -7$, $z = -14$; anterior STG/S: $x = 57$, $y = 8$, $z = -11$. All ROI masks were created using SPM12.

3.2.2 | Contrasts of interest

We defined two different contrasts of interest to test whether BOLD responses increased in the visual ROIs during voice-identity recognition for voice-face learned speakers (in contrast to voice-occupation learned speakers) in high, compared to low, noise listening conditions. First, we calculated contrast maps (t-statistics), examining our central hypothesis, for the interaction term: ([voice-face/high-noise > voice-occupation/high-noise] > [voice-face/low-noise > voice-occupation/low-noise]). Second, we performed post-hoc t-tests by investigating the effect of learning (voice-face vs. voice-occupation) on responses in face-sensitive regions of the FFA and pSTS-mFA separately for both high- and low-noise listening conditions. Here, we calculated contrast maps (t-statistics) for the two simple main effects of learning for: high-noise listening conditions (voice-face/high-noise > voice-occupation/high-noise) and low-noise listening conditions (voice-face/low-noise > voice-occupation/low-noise). Responses in each ROI were considered to be significant if they were present at p <.05 family wise error (FWE) corrected for the ROI, Holm–Bonferroni (Holm, 1979) corrected for the number of ROIs ($N = 3$). The Holm–Bonferroni method handles multiple comparisons via a sequential hypothesis rejection approach, and it is less susceptible to Type II error (i.e., missing true effects) compared to the standard single-step Bonferroni correction (Nichols & Hayasaka, 2003).

3.2.3 | Correlational analyses

For testing whether the magnitude of responses in face-sensitive regions during the voice-identity recognition task correlated with behavioural face-benefit scores across participants we performed the following steps. First, we calculated the behavioural face-benefit score (von Kriegstein et al., 2008): % correct voice-identity recognition for voice-face learning minus % correct voice-identity recognition for -voice-occupation learning. This score was calculated separately for high- and low-noise listening conditions, for each participant. We included the face-benefit score for high- and low-noise as a co-variate of interest in SPM12 for the second-level analysis of the simple main effects of learning for high- (voice-face/high-noise > voice-occupation/high-noise) and low- (voice-face/low-noise > voice-occupation/low-noise) noise listening conditions, respectively.

Secondly, as we noted in our previous behavioural study that the face-benefit increased in higher noise-levels (i.e., decreasing SNRs; Maguinness et al., 2021), we calculated an additional score which reflected how well the behavioural face-benefit was maintained in the high-, relative to the low-, noise listening conditions (face-benefit high-noise minus face-benefit low-noise). We refer to this score as 'face-benefit maintenance'. This score was calculated for each participant and was included in SPM12 as a co-variate of interest at the second-level analysis for the interaction contrast ([voice-face/high-noise > voice-occupation/high-noise] – [voice-face/low-noise > voice-occupation/low-noise]).

The significance of the correlational analyses was assessed using SPM12 and considered to be significant if present at p <.05 FWE corrected for the ROI, Holm–Bonferroni-corrected for the number of ROIs ($N = 3$).

3.2.4 | Psychophysiological interactions analysis

To test whether there is functional connectivity (Friston, 1994), between visual face- and voice-sensitive regions during voice-identity recognition in high-noise, we conducted psychophysiological interactions (PPI) analyses (Friston et al., 1997; O'Reilly, Woolrich, Behrens, Smith, & Johansen-Berg, 2012). PPI analyses identify temporal correlations between responses in specific brain regions (i.e., seed regions) and responses in other brain regions (i.e., target regions) which are modulated by a psychological factor (i.e., experimental task). The seed regions for the PPI analysis were the visual face-sensitive regions (as defined by the visual face area localiser) which demonstrated selective enhanced responses for face-learned speakers in high-noise listening conditions ([voice-face/high-noise > voice-occupation/high-noise] > [voice-face/low-noise > voice-occupation/low-noise]). Target regions were the voice-sensitive mid and anterior STG/S.

We conducted the PPI analyses following standard procedures (Friston et al., 1997). We extracted the first Eigenvariate in the visual face-sensitive seed region and used the voice-sensitive regions in the mid and anterior STG/S regions as target regions. The analysis included the psychological variable for the simple main effects of learning for the high-noise listening condition (voice-face/high-noise > voice-occupation/high-noise). We modelled the first Eigenvariate, the psychological variable, and the psychophysiological interaction term as regressors at the single-subject level. The psychophysiological interaction term was created using routine procedures implemented in SPM12. Population-level inferences about BOLD signal changes were based on a random-effects model that estimated the second-level statistic at each voxel using a one-sample *t*-test. The face-benefit score for each participant, for the high-noise condition, was included as a co-variate of interest in the second-level analysis. Results were considered significant if they were present at *p* <.05 FWE corrected for the target ROI, Holm–Bonferroni corrected for the number of target ROIs ($N = 2$).

4 | RESULTS

4.1 | Behavioural results: auditory-only voice-identity recognition

4.1.1 | Accuracy

The 2 × 2 repeated measures analysis of variance (ANOVA), with ‘learning’ (voice-face or voice-occupation) and ‘noise-level’ (high-noise or low-noise) as repeated factors on the accuracy scores (Table 1) revealed a main effect of ‘learning’ ($F_{(1,20)} = 5.91$, $p = .02$, $\eta_p{}^2 = .23$). As expected, this main effect was based on higher accuracy for recognising the identity of speakers who had been previously learned through voice-face ($M = 84.9\%$), rather than voice-occupation ($M = 78.6\%$), learning. We refer to such an improvement in performance as the ‘face-benefit’ (von Kriegstein et al., 2008); 16 of the 21 participants showed this average effect across noise levels (Figure 5). A main effect of ‘noise-level’ was also observed [$F_{(1,20)} = 37.75$, $p = <.001$, $\eta_p{}^2 = .65$], with lower recognition accuracy in high-noise ($M = 77.9\%$), compared to low-noise ($M = 85.6\%$), listening conditions. Contrary to our expectations, there was no significant interaction between ‘learning’ and ‘noise-level’ [$F_{(1,20)} = 2.47$, $p = .13$, $\eta_p{}^2 = .10$], suggesting that the difference in the face-benefit across noise conditions was not significant. The face-benefit in the high-noise condition was 5.4%, and in the low-noise condition was 7.3%. The face-benefit for high- and low-noise conditions was strongly positively correlated within participants (Pearson's $r = .909$, $p = .000$, $N = 21$; Pearson's $r = .847$, $p = .000$, $N = 16$ positive face-benefit participants only; Figure 5).

TABLE 1 Behavioural accuracy and reaction times for auditory-only voice-identity recognition performance (with standard deviations) for voice-face and voice-occupation learned speakers, in high- and low-noise listening conditions

	High-noise	Low-noise
Voice-face		
% correct	80.6 (7.5)	89.3 (8.6)
Reaction time (ms)	612.1 (54.7)	604.9 (55.6)
Voice-occupation		
% correct	75.1 (11.1)	82.0 (9.6)
Reaction time (ms)	625.1 (60.5)	622.5 (67.4)
Face-benefit		
% correct	5.4 (12.9)	7.3 (11.6)
Reaction time (ms)[a]	−12.9 (32.6)	−17.6 (40.7)

Abbreviation: ms, milliseconds.

[a]Negative values indicate comparatively faster responses.

4.1.2 | Reaction time

A 2 × 2 repeated measures analysis of variance (ANOVA) was conducted on the time taken to recognise speaker identities across ‘learning’ (voice-face or voice-occupation) and ‘noise-level’ (high-noise or low-noise) conditions (Table 1). A main effect of ‘learning’ was found ($F_{(1,20)} = 4.36$, $p = .05$, $\eta_p{}^2 = .18$). Participants were faster to recognise the identities of speakers who had been previously learned through voice-face ($M = 608$ ms), compared to voice-occupation ($M = 624$ ms), learning. This indicates that there was no speed accuracy trade-off for the main effect of learning. The main effect of ‘noise-level’ ($F_{(1,20)} = 1.73$, $p = .20$, $\eta_p{}^2 = .08$) and the interaction

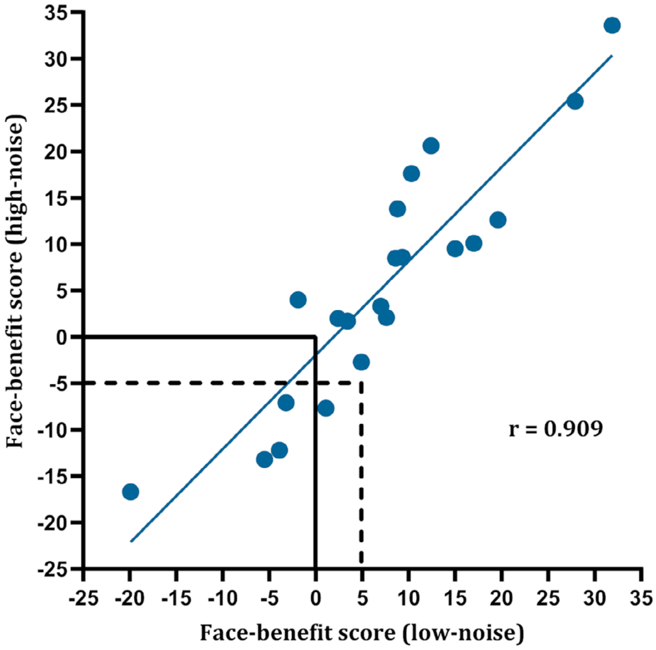

FIGURE 5 Plot showing the behavioural face-benefit score for each participant across high- (*y* axis) and low-noise (*x* axis) listening conditions. There was a significant positive correlation ($r = .909$, $p = .000$) between participants' face-benefit scores across noise levels. The bold black intersecting line denotes no face-benefit in either noise level (i.e., accuracy is equal for voice-face and voice-occupation learned speakers). The dashed intersecting line denotes a division between those with a positive ($N = 16$) and negative ($N = 5$) face-benefit score averaged over both noise levels

between 'learning' and 'noise-level' ($F_{(1,20)} = 0.49$, $p = .49$, $\eta_p^2 = .02$) were not significant.

4.2 | Functional MRI results

4.2.1 | Increased responses in the right pSTS-mFA during the recognition of face-learned speakers in high-noise

To address our main hypothesis, we examined whether the recognition of face-learned speakers in higher versus lower noise was associated with increased responses in the FFA and/or the pSTS-mFA. To do this, we used the interaction contrast ([voice-face/high-noise > voice-occupation/high-noise] > [voice-face/low-noise > voice-occupation/low-noise]). For this contrast, we observed increased responses in the right pSTS-mFA ($x = 51$, $y = -37$, $z = 8$, $p = .012$, FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs; Figure 6a,b). There were also increased responses in the left pSTS-mFA, although this did not survive Holm–Bonferroni-correction ($x = -54$, $y = -43$, $z = 8$, $p = .027$, FWE corrected for the ROI; Figure S2 Supporting Information). Contrary to our expectations, no increased responses were observed in the FFA for the interaction contrast even at a lenient threshold (p <.05, uncorrected). To confirm the directionality

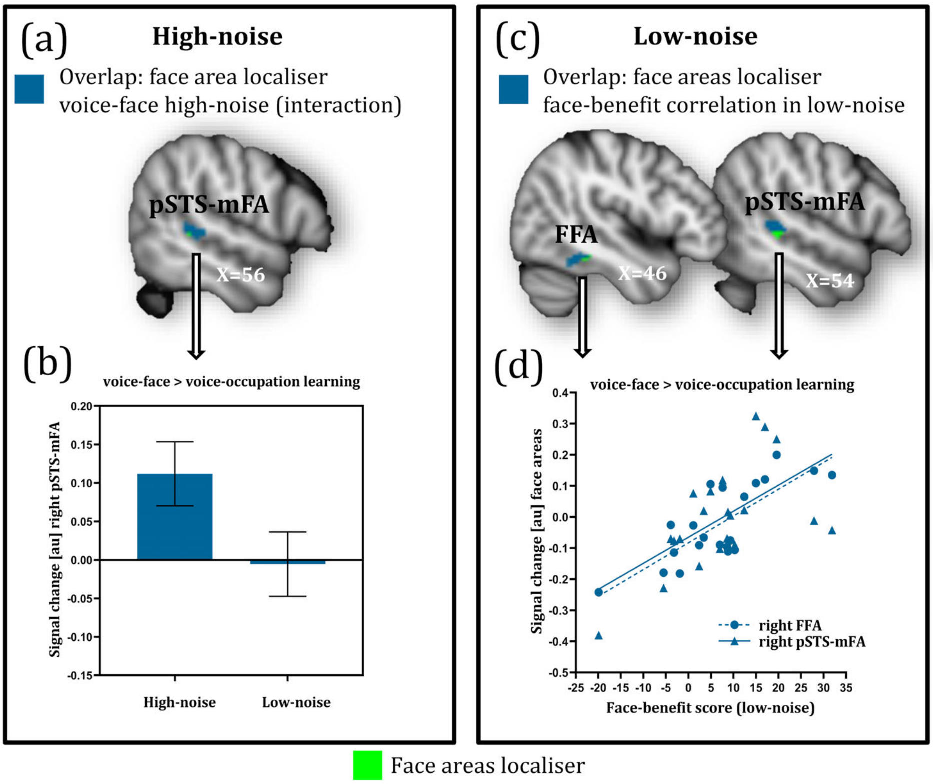

FIGURE 6 fMRI results. Overview of responses in visual face areas (FFA, pSTS-mFA) during voice-identity recognition for face-learned speakers in different levels of auditory noise. (a) and (b) High-noise. (a) Overlay of the interaction contrast ([voice-face high-noise > voice-occupation high-noise] > [voice-face low-noise > voice-occupation low-noise]) and the functional face area localiser contrast for the right pSTS-mFA (dynamic faces > dynamic objects). The overlap between the contrasts is shown in blue. The maximum statistic for the interaction contrast was at $x = 51$, $y = -37$, $z = 8$, $p = .012$, FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs. (b) Signal change estimate in the right pSTS-mFA for conditions that were included in the interaction contrast ([voice-face high-noise > voice-occupation high-noise] > [voice-face low-noise > voice-occupation low-noise]) during auditory-only voice-identity recognition. Plot shows first eigenvariate extracted from the maximum statistic for the interaction contrast shown in (a). (c) and (d) Low-noise. (c) Overlay of the correlation between the face-benefit score for each participant and the contrast (voice-face low-noise > voice-occupation low-noise) and the functional localiser contrasts for the right FFA (faces > objects) and right pSTS-mFA (dynamic faces > dynamic objects) (overlap between the contrasts is shown in blue). (d) Plot of the correlation between the face-benefit score and responses in right FFA and right pSTS-mFA for the contrast (voice-face > voice-occupation) in low-noise listening conditions. The signal change estimate (first eigenvariate) was extracted from the peak co-ordinates for the right FFA ($x = 48$, $y = -40$, $z = -16$, $p = .000$, FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs) and the right pSTS-mFA ($x = 57$, $y = -31$, $z = 5$, $p = .021$, FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs) for the contrast (voice-face > voice-occupation) in low-noise, with the face-benefit score (low-noise) as a co-variate of interest. The plot is shown for display purposes only. Statistics were computed in SPM12 (see Section 3). In (a) and (c) the overlapping responses (blue colours) between the voice-identity recognition contrasts and the face areas are presented for display purposes at $p = .05$ whole brain uncorrected, masked by the ROI. Non-overlapping responses in visual face areas are shown in green. All responses are overlaid on a mean structural MNI152 T1 weighted image (sagittal view, right hemisphere) and visualised using MRIcron (www.nitrc.org/projects/mricron). All co-ordinates are reported in MNI-space

of the interaction we examined responses in the visual ROIs for face-learned speakers separately for high- and low-noise listening conditions. The tests of simple main effects of learning confirmed increased responses in the right pSTS-mFA ($x = 51$, $y = -34$, $z = 1$, $p = .028$, FWE corrected for the ROI) for the contrast (voice-face/high-noise > voice-occupation/high-noise). In contrast, there was no evidence for increased responses in the left or right pSTS-mFA or FFA for the contrast (voice-face/low-noise > voice-occupation/low-noise), even at lenient thresholds (p <.05, uncorrected).

We noted, in line with previous findings (Maguinness et al., 2021; von Kriegstein et al., 2008), that not all participants had a behavioural face-benefit (Figure 5). To further examine the functional relevance of responses in the visual ROIs for the face-benefit on speaker recognition, we separately examined responses for the 16 participants who showed this behavioural enhancement. The responses in the visual ROIs (Table S1 Supporting Information, ROIs for $N = 15$) stayed qualitatively the same: there was a noise-level × learning interaction in the right pSTS-mFA ($x = 54$, $y = -28$, $z = -1$, $p = .030$, FWE corrected for ROI), which did, however, not survive Holm–Bonferroni correction. The left pSTS-mFA was not significant ($x = -54$, $y = -43$, $z = 8$, $p = .097$, FWE corrected for ROI). Tests of simple main effects of learning showed significantly increased responses in the right pSTS-mFA ($x = 51$, $y = -31$, $z = -1$, $p = .004$, FWE corrected for the ROI, Holm–Bonferroni corrected for number of ROIs), for the contrast (voice-face/high-noise > voice-occupation/high-noise). Responses for this contrast in the left pSTS-mFA ($x = -51$, $y = -43$, $z = 8$, $p = .042$, FWE corrected for the ROI) did not survive Holm–Bonferroni correction. No increased responses were observed in the left or right pSTS-mFA or FFA for the contrast 'voice-face/low-noise > voice-occupation/low-noise', even at lenient thresholds (p <.05, uncorrected). Taken together, these findings suggest that motion-sensitive regions of the face-network, particularly in the right hemisphere, may be engaged for voice-identity recognition in high-noise listening conditions.

4.2.2 | The face-benefit is positively correlated with increased functional responses in the FFA and the right pSTS-mFA in low-noise listening conditions

Previous studies have demonstrated responses in the FFA during the recognition of speakers known by face (Blank et al., 2011; Schall et al., 2013; von Kriegstein et al., 2005; von Kriegstein et al., 2006; von Kriegstein & Giraud, 2006). Pertinently, these responses have been shown to be behaviourally relevant for supporting voice-identity recognition: They correlated positively with the face-benefit score (von Kriegstein et al., 2008) in typically developed individuals, but not in developmental prosopagnosics (individuals with a severe deficit in face-identity processing; McConachie, 1976) who do not have a face-benefit (von Kriegstein et al., 2006; von Kriegstein et al., 2008). The categorical results from the previous section suggest that the pSTS-mFA may be involved in voice-identity recognition in more noisy listening conditions, although the behavioural relevance of these responses remains unclear. Therefore, we examined a possible correlation between the behavioural face-benefit and responses in the visual ROIs for both the high- (voice-face/high-noise > voice-occupation/ high-noise) and low- (voice-face/low-noise > voice-occupation/low-noise) noise listening conditions. Contrary to our expectations, in high-noise listening conditions, the face-benefit (calculated for the high-noise condition) did not correlate with responses in *any* visual ROIs for the contrast 'voice-face/high-noise > voice-occupation/high-noise'. In low-noise listening conditions, we found a significant positive correlation between the face-benefit (calculated for the low-noise condition) and responses in the FFA ($x = 48$, $y = -40$, $z = -16$; $p = .000$; FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs). There was also a positive correlation between the behavioural face-benefit and responses in the right pSTS-mFA in the - low-noise condition ($x = 57$, $y = -31$, $z = 5$, $p = .021$; FWE corrected for the ROI, Holm–Bonferroni corrected for number of ROIs; Figure 6c,d). For the analyses with the 16 participants with a face-benefit the low-noise listening condition results for the FFA remained ($p = .006$; FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs). However, the pSTS-mFA (left or right) did not show a significant correlation in this sub-group (p >.05, uncorrected).

4.2.3 | Are responses in the right pSTS-mFA associated with the ability to maintain the face-benefit in high-noise listening conditions?

It was surprising that in high-noise conditions there were no positive correlations between BOLD responses in the pSTS-mFA and the face-benefit score. This could mean that the recruitment of pSTS-mFA during high-noise reflects an attempt to compensate, but ultimately fails to be of behavioural relevance for supporting voice-identity recognition. Previously, we observed that for participants who have a face-benefit (76% of participants—Maguinness et al., 2021), there is a linear increase in the face-benefit with increasing auditory noise. Contrary to this, we did not observe a larger face-benefit in the high-noise condition of the present study (see Table 1). However, we noted variability in participants behavioural performance—not all participants maintained the same degree of face-benefit across noise levels. For some this benefit dropped substantially in the high-, compared to low-noise, condition. While for others the face-benefit was equatable or even greater in high-noise (Figure 5). An alternative view may therefore be that the pSTS-mFA may be behaviourally relevant for *maintaining* the face-benefit in noise. To that end, we conducted an exploratory analysis which examined the correlation between the 'face-benefit maintenance' score (face-benefit high-noise minus face-benefit low-noise) and the interaction contrast ([voice-face/high-noise > voice-occupation/high-noise] > [voice-face/low-noise > voice-occupation/low-noise]) in the visual ROIs. The correlation was not statistically significant in the right pSTS-mFA ($x = 63$, $y = -43$, $z = 8$; $p = .069$, FWE corrected for the ROI) or the other two ROIs even at lenient thresholds (p <.05, uncorrected). When the analysis included only the 16 participants who had demonstrated a face-benefit on speaker recognition there was a positive correlation between the face-benefit maintenance score and responses in the right pSTS-mFA (*x* = 48, *y* = −34, *z* = −4; *p* = .024, FWE corrected for the ROI), although this did not survive Holm–Bonferroni correction for the three ROIs (Figure 7). Neither the left pSTS-mFA nor the FFA correlated with this measure in this sub-group (*p* >.05, uncorrected). This finding suggests a potential behaviourally relevant role for the right pSTS-mFA in preserving the beneficial effect of face experience on voice-identity recognition in high-levels of auditory noise, for participants who benefit from audio-visual voice-face learning. However, given that it did not survive Holm–Bonferroni correction and was an exploratory analysis it must be taken with caution.

4.2.4 | Functional connectivity between pSTS-mFA and voice-sensitive regions in the right STG/S is associated with the face-benefit in high-noise listening conditions

In previous work it was shown that the FFA is functionally coupled with voice-sensitive regions in the anterior and mid STG/S during voice-identity recognition of face-learned speakers (Schall & von Kriegstein, 2014; von Kriegstein et al., 2005; von Kriegstein et al., 2006; von Kriegstein & Giraud, 2006). Based on these findings we expected that the right pSTS-mFA, associated with the recognition of face-learned speakers in high-noise (Figure 6a,b), would also share functional connections with voice-sensitive regions in the STG/S. To test whether the pSTS-mFA is functionally connected to anterior and mid STG/S voice-sensitive regions, we conducted PPI analyses, including the psychological variable (voice-face high-noise > voice-occupation high-noise). We defined the right pSTS-mFA as the seed region and the voice-sensitive mid STG/S and anterior STG/S as target regions. There were no significant results in any of the target ROIs, even at a lenient threshold of *p* <.05 uncorrected, for analyses including *N* = 21, or *N* = 16 face-benefit participants.

Next, we explored whether there might be a correlation between functional connectivity of face- and voice-sensitive regions in the STG/S during voice-identity recognition in high-noise listening conditions and the amount of the face-benefit across participants. The analysis included the psychological variable (voice-face high-noise > voice-occupation high-noise) and the behavioural covariate (face-benefit high-noise) for all participants (*N* = 21). The correlation for all *N* = 21 between the face-benefit and increased functional connectivity between the right pSTS-mFA (seed region) and either voice-sensitive regions in the mid or anterior STG/S was not statistically significant (mid STG/S at *x* = 63, *y* = −4, *x* = −10, *p* = .064, FWE corrected for the ROI; anterior STG/S at *x* = 60, *y* = 2, *x* = −7, *p* = .098, FWE corrected for the ROI). Interestingly, the correlation was only significant for the 16 participants who displayed a face-benefit on voice-identity recognition: This was the case for both the connectivity between the right pSTS-mFA and the mid voice-sensitive STG/S (*x* = 63, *y* = −1, *z* = −10; *p* = .008, FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs) and the anterior voice-sensitive STG/S (*x* = 60, *y* = 2, *z* = −7; *p* = .024, FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs; Figure 8; Table 2).

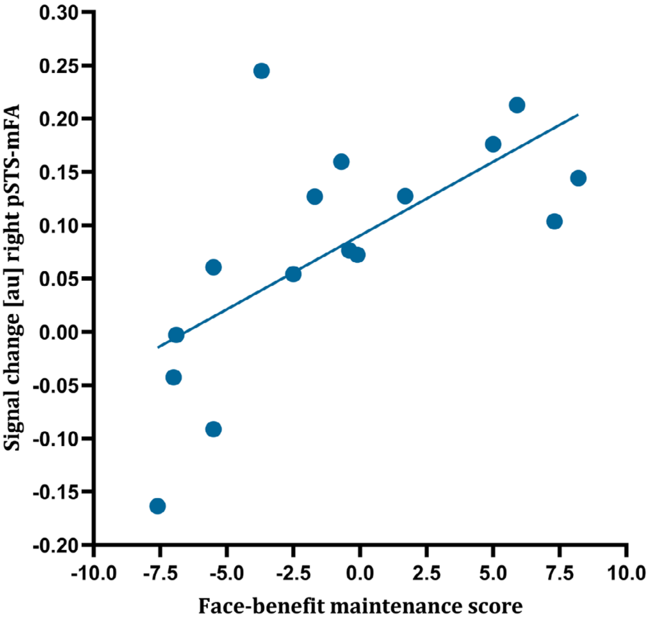

FIGURE 7 Plot showing the correlation between the face-benefit maintenance score (*N* = 16 face-benefit participants) and functional responses in the right pSTS-mFA (first eigenvariate extracted from peak co-ordinate *x* = 48, *y* = −34, *z* = −4, *p* = .024, FWE corrected for ROI) for the interaction contrast ([voice-face high-noise > voice-occupation high-noise] > [voice-face low-noise > voice-occupation low-noise]). The correlation was computed in SPM12 and the plot serves for display purposes, only. The correlation did not survive Holm–Bonferroni correction for three ROIs and therefore should be interpreted with caution

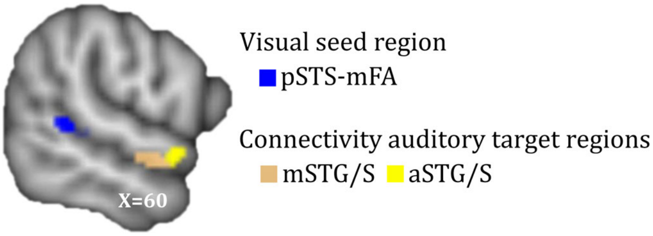

FIGURE 8 Functional connectivity between face- and voice-sensitive regions during voice-identity recognition for face-, in comparison to occupation-, learned speakers in high-noise. In high-noise listening conditions, in participants who benefitted from voice-face learning (*N* = 16), the face-benefit correlated positively with increased connectivity between the right pSTS-mFA (blue; dynamic faces > dynamic objects) and auditory voice-sensitive regions in the mid and anterior STG/S (beige/yellow). For display purposes, results for auditory target regions are shown at *p* = .05 whole brain uncorrected, masked by the 8 mm spheres centred on previously published co-ordinates for the mid and anterior STG/S (Blank et al., 2011)

TABLE 2 Peak co-ordinates for voice-sensitive regions which showed increased functional connectivity to the right pSTS-mFA during the recognition of voice-face learned, in contrast to voice-occupation learned, speakers in high-noise listening conditions

Seed region: right pSTS-mFA								
	x	*y*	*z*	*T*-value	*x*	*y*	*z*	*T*-value
	All participants ($N = 21$)				Face-benefit participants ($N = 16$)			
Right mSTG/S	*63*	*−4*	*−10*	*2.64*	**63**	**−1**	**−10**	**4.64**
Right aSTG/S	*60*	*2*	*−7*	*2.66*	**60**	**2**	**−7**	**3.85**

Note: Co-ordinates in bold are significant for the face-benefit participants ($N = 16$, right mSTG/S $p = .008$, aSTG/S $p = .024$, FWE corrected for ROI, Holm–Bonferroni corrected for number of ROIs), co-ordinates in italics are non-significant ($N = 21$, all participants) and are included for comparison purposes only.

5 | DISCUSSION

We investigated whether the visual mechanisms used during the recognition of auditory-only communication signals are adaptive to different levels of noise in the auditory signal. Our study had two key findings. First, and centrally, in high-noise listening conditions we observed that the recognition of face-learned speakers recruited the motion-sensitive right pSTS-mFA. Unexpectedly, responses in this region did not correlate with listeners' behavioural face-benefit scores. Only the functional connectivity between the right pSTS-mFA and voice-sensitive regions correlated with the behavioural face-benefit in the high-noise listening condition, in the 16 participants with a face-benefit. Conversely, in low-noise, there was no categorical response to face-learned, in contrast to occupation learned, speakers in any visual face-sensitive ROI. However, in this noise level, the behavioural face-benefit was robustly correlated with increased functional responses in the region sensitive to structural facial-identity cues i.e., the FFA and—to some extent—with the right pSTS-mFA. The findings suggest that partially distinct visual mechanisms support the -face-benefit in different levels of auditory noise.

5.1 | Visual responses during voice-identity recognition in noise: support for an audio-visual model of human auditory communication

The visual pSTS-mFA has been implicated in the processing of dynamic facial cues, including those dynamic cues which support identity processing (Girges et al., 2015, 2016; O'Toole et al., 2002). Our findings suggest that voice-identity recognition in high-noise, when listeners arguably attend to more dynamic aspects of the voice for recognition, may stimulate the engagement of stored dynamic, rather than static, identity cues encoded during audio-visual voice-face learning. Such a finding corroborates previous observations that voice-identity recognition is facilitated by dynamic identity cues available in the auditory and visual streams. For example, Schweinberger and Robertson (2017) noted that familiar voices were more readily recognised when accompanied with their time synchronised face (i.e., video), while recognition was impaired when the corresponding moving face was presented with a temporal asynchrony. The present findings imply that even in the absence of the concurrent face input, the recruitment of a brain region implicated in dynamic face processing may support voice-identity processing, particularly when static auditory cues are degraded.

Conversely, in low-noise listening conditions, when recognition may rely on more available static vocal properties (Figure 2), we noted a robust relationship between the face-benefit and responses in the FFA. Responses in this facial form-sensitive region correlated with individuals' recognition scores for face, in comparison to non-face, learned speakers. While we found a similar correlation with responses in the right pSTS-mFA, this correlation did not survive for individuals who displayed a positive face-benefit score. Together, these findings of responses in different face-sensitive regions, modulated by noise in the auditory signal, highlight the remarkably adaptive nature of cross-modal responses observed under unisensory listening conditions. They support central assumptions of the audio-visual model of human auditory communication (von Kriegstein, 2012) by demonstrating that: (a) the visual regions engaged during auditory-only processing may be linked to the common source, task relevant, identity information available in both modalities (i.e., dynamic or static cues); (b) persistent responses in visual regions during noisy listening conditions serve to enhance the processing of voices, rather than being an epiphenomenon of successful speaker-identity recognition (Schall et al., 2013; von Kriegstein et al., 2006).

5.2 | The face-benefit across noise-levels

Based on previous behavioural findings (Maguinness et al., 2021) we had expected that the face-benefit would be greatest in the high-noise condition. Surprisingly, this was not the case. There are two likely reasons for the lack of difference across noise levels. One may relate to the number of SNRs tested. Our previous observation of a linear increase in the face-benefit with increasing noise (Maguinness et al., 2021) tested a greater number of SNRs. Potentially, the linear effect may be more apparent with this larger range. However, the low-noise condition in the present study was associated with a quantitatively (but not significantly) higher face-benefit than the high-noise condition. Another possibility is therefore that the face-benefit increases linearly with decreasing SNRs (Maguinness et al., 2021), but starts to break down at a certain point where the SNR is so low that static, and potentially also dynamic cues, cannot be tracked reliably.

Although our previous study showed the linear effect up to SNR −8 dB, in the present study the more challenging listening conditions in the MRI-environment could have impacted the audibility of the high-noise level.

5.3 | Relevance of the pSTS-mFA for the face-benefit

Unlike the FFA, responses in the right pSTS-mFA, particularly in high-noise listening conditions, did not correlate directly with the (positive) face-benefit score. These findings might question the behavioural relevance of responses in the pSTS-mFA. While this lack of correlation may relate to the saliency of the dynamic cues available in the SNRs tested, it is additionally possible that it may also relate to the difference in the time courses for the acquisition of structural form and dynamic identity cues. For example, although dynamic facial identity cues can be learned, it has been suggested that they become more robust with repeated exposure and the degree of idiosyncrasy (Butcher & Lander, 2017; Lander & Chuang, 2005; O'Toole et al., 2002; Roark et al., 2003). Thus, during our audio-visual training the acquired dynamic identity signature may have been less robust, than its corresponding more rapidly acquired structural identity representation (Blank et al., 2015). This may possibly explain why the face-benefit correlated directly with FFA responses, but not with responses in the pSTS-mFA. Potentially, extending the audio-visual training period, or adding auditory noise during training, may enhance the acquisition of these dynamic cues. An alternative explanation is that responses in the pSTS-mFA may not be as stable at supporting identity processing compared to structural cues in the FFA. For example, while individuals with developmental prosopagnosia can use dynamic cues to recognise faces in laboratory settings (Longmore & Tree, 2013; Steede, Tree, & Hole, 2007), they nevertheless fail to recognise faces in day-to-day interactions, highlighting that dynamic cues alone may not be sufficient to support typical identity processing (Maguinness & Newell, 2015). Notwithstanding these considerations, the connectivity results from the main fMRI experiment support behaviourally relevant cross-modal interactions between dynamic face and voice regions: as the face-benefit (in participants who benefitted for voice-face learning) in high-noise correlated with increased functional connectivity between the pSTS-mFA and the voice-sensitive STG/S. We take this behaviourally relevant connectivity profile as first evidence that dynamic face cues and processing in the pSTS-mFA may support the auditory-only processing of face-learned speakers, particularly in degraded listening conditions. Interestingly, functional connections between the *left* face-sensitive pSTS and a speech intelligibility region in the left anterior STS have been shown to support speech recognition for face learned speakers (Schall & von Kriegstein, 2014). While we corroborate a similar AV network, we demonstrate that for voice-identity recognition it is likely associated with responses in the right pSTS. This connectivity profile is in line with the right hemisphere's dominant role in identity processing (Assal, Zander, Kremin, & Buttet, 1976; Barton, 2008; Belin & Zatorre, 2003; De Renzi, 1986; Kanwisher et al., 1997; Liu, Corrow, Pancaroglu, Duchaine, & Barton, 2015; Luzzi et al., 2018; Rossion, 2014; von Kriegstein & Giraud, 2004).

5.4 | An audio-visual voice-face network along the STS for voice-identity processing

Recently, Yovel and O'Toole (2016) proposed that recognition of the 'dynamic speaking person' was likely mediated solely by voice and face processing regions along the STS which are sensitive to *temporal* information and dismissed a potential role for interactions with the FFA. Importantly, while we documented evidence of a motion-sensitive AV network we demonstrate that it is likely complementary, rather than fundamental, for supporting voice-identity recognition. In a similar vein to face-identity recognition, the network appears to be recruited as a complementary, potentially 'back-up', system for supporting voice-identity recognition when static cues are altered or unavailable. We propose that the AV voice-face network along the STS might systematically supplement the FFA mechanism, that is, becoming increasingly more responsive, as static aspects of the auditory signal are degraded. This is suggested by our finding of behaviourally relevant responses in both the FFA and to a lesser degree in the right pSTS-mFA during voice-identity recognition in low-noise. Conversely, in high-noise the recognition of face-learned speakers engaged the pSTS-mFA *only*. Such a system is in line with the visual literature which has demonstrated that the perceptual system integrates both static and dynamic face-identity cues in a manner which is dependent on either cues perceived saliency (Dobs, Ma, & Reddy, 2017; Knappmeyer, Thornton, & Bülthoff, 2003). While Yovel and O'Toole (2016) and other AV models of person-identity processing are mostly tailored towards the brain mechanisms supporting audio-visual *integration* that is, when both the face and voice are concurrently presented (e.g., Young, Frühholz, & Schweinberger, 2020), our findings nevertheless highlight the importance of considering how both dynamic *and* static AV identity cues might be integrated. Given the perceptual system's sensitivity to static and dynamic components in the AV person-identity signal, we deem it is unlikely that integration is governed solely by a common global mechanism in the STS.

5.5 | The pSTS as a multimodal region and voice-sensitive region

The posterior STS has been associated with the multimodal representation and integration of faces and voices (Tsantani, Kriegeskorte, McGettigan, & Garrido, 2019; Young et al., 2020), compared to for example, objects and sounds (Watson, Latinus, Charest, Crabbe, & Belin, 2014). Thus, the region could be conceived as an audio-visual person-identity representational hub, rather than a region which is sensitive to different aspects of identity including facial dynamics. Theoretically, it is possible that a multimodal pSTS may be additionally recruited to support the recognition of face-learned voices in noise. However, two points speak against this. First, we functionally localised the motion-sensitive face area in the pSTS with a specific localiser contrasting dynamic faces against dynamic objects (group [$N = 20$] peak voxel location: $x = 54$, $y = -34$, $z = 2$) and localised the peak responses for face-learned speakers in high-noise at $x = 51$, $y = -37$, $z = 8$ (group ($N = 21$) peak voxel location; noise-level × learning interaction). The peak response for the interaction was >1 cm away from those reported for multimodal voice-face representations ($x = 48$, $y = -49$, $z = 11$; searchlight analysis for cross-modal generalisation of discriminants for pairs of identities [Tsantani et al., 2019]). Second, if the observed pSTS responses were an additional audio-visual integrative mechanism recruited to complement the FFA, we would expect FFA responses to be equally present for the high-noise conditions. This was not the case.

Additionally, the pSTS has been implicated in voice-identity processing, particularly for unfamiliar voices which require increased perceptual processing (Schelinski, Borowiak, & von Kriegstein, 2016; von Kriegstein & Giraud, 2004). Thus, it could be argued that the observed enhanced pSTS responses in the current study may have been driven solely by increased voice-identity processing in more challenging listening conditions. However, if this were the case, we would expect an overall (i.e., regardless of learning condition) response increase in this region during voice-identity processing in high-, compared to low-, noise listening conditions. This was not evident. In contrast, the pSTS-mFA responses were observed specifically for face-learned speakers in noise (i.e., interaction effect) and not as a main effect for processing voices in noisier listening conditions (see Supporting Information, Functional MRI Analysis).

5.6 | Interindividual variability in the face-benefit

We noted variability in how well participants maintained the face-benefit in high-, compared to, low-noise listening conditions. Based on an exploratory analysis, there were some indications that this variability may relate to responses in the right pSTS-mFA, such that higher face-benefit maintenance scores were correlated with increased functional responses within this region. However, it is important to note that this correlation analysis was exploratory and did not survive Holm–Bonferroni correction and should be interpreted with caution. This observation was restricted to the 16 individuals who benefitted from face-voice learning, that is, 76% of the tested sample. Currently it is unclear why some individuals do not benefit from face-voice learning. Although findings from developmental prosopagnosia (McConachie, 1976), that is, a severe deficit in face-identity processing, suggest that it may be related to face processing abilities (Maguinness & von Kriegstein, 2017; von Kriegstein et al., 2006; von Kriegstein et al., 2008). Other evidence of interactions (Bülthoff & Newell, 2015, 2017) and relationships between face- and voice-identity recognition abilities in the neurotypical population (Jenkins et al., 2020), suggest that a common coding system may underpin this enhancement i.e., similar computations in different modalities. Interestingly, the proportion of the current sample with a face-benefit is in line with our previous observations. For example, von Kriegstein et al. (2008) observed a face-benefit for voice-identity recognition in 13 of the 17 participants tested. While in Maguinness et al. (2021) this face-benefit was present in 19 of 25 individuals.

6 | CONCLUSION

In summary, we propose that during audio-visual learning a vocal identity becomes enriched with distinct visual features, pertaining to both static and dynamic aspects of facial identity. These stored visual cues are used in an adaptable manner, tailored to perceptual demands, to optimise subsequent auditory-only voice-identity recognition. In more optimal listening conditions, the FFA is recruited to enhance voice-identity recognition. In contrast, under more degraded listening conditions, the facial motion-sensitive pSTS-mFA is recruited, although this complementary mechanism may be potentially less beneficial for supporting voice-identity recognition than that of the FFA. Taken together, these findings corroborate and extend an audio-visual view of human auditory communication, providing evidence for the particularly adaptive nature of cross-modal responses and interactions observed under unisensory listening conditions.

ACKNOWLEDGMENTS

The authors thank Kamila Borowiak for comments on an earlier version of this manuscript, Sonja Schall for auditory and visual stimuli, and Stefanie Schelinski for providing the visual-only face area localiser experiment. This work was funded by a Max Planck Research Group grant to Katharina von Kriegstein. Corrina Maguinness also receives funding from the ERC-consolidator grant SENSOCOM 647051 and the DFG KR 3735/5-1.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

AUTHOR CONTRIBUTIONS

Corrina Maguinness and Katharina von Kriegstein designed the research; Corrina Maguinness performed the research and data analysis; Corrina Maguinness and Katharina von Kriegstein wrote the article.

DATA AVAILABILITY STATEMENT

Data Availability: The data that support the findings of this study are available upon reasonable request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Corrina Maguinness https://orcid.org/0000-0002-6200-4109
Katharina von Kriegstein https://orcid.org/0000-0001-7989-5860

REFERENCES

Assal, G., Zander, E., Kremin, H., & Buttet, J. (1976). Discrimination des voix lors des lesions du cortex cerebral. *Archives Suisses de Neurologie, Neurochirurgie et de Psychiatrie*, *119*(2), 307–315 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1006205

Axelrod, V., & Yovel, G. (2015). Successful decoding of famous faces in the fusiform face area. *PLoS One*, *10*(2), e0117126. https://doi.org/10.1371/journal.pone.0117126

Barton, J. J. S. (2008). Structure and function in acquired prosopagnosia: Lessons from a series of 10 patients with brain damage. *Journal of Neuropsychology*, *2*(1), 197–225. https://doi.org/10.1348/174866407X214172

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport*, *14*(16), 2105–2109. https://doi.org/10.1097/01.wnr.0000091689.94870.85

Bernstein, M., Erez, Y., Blank, I., & Yovel, G. (2018). An integrated neural framework for dynamic and static face processing. *Scientific Reports*, *8*(1), 2–11. https://doi.org/10.1038/s41598-018-25405-9

Bernstein, M., & Yovel, G. (2015). Two neural pathways of face processing: A critical evaluation of current models. *Neuroscience and Biobehavioral Reviews*, *55*, 536–546. https://doi.org/10.1016/j.neubiorev.2015.06.010

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice- and face-recognition areas. *The Journal of Neuroscience*, *31*(36), 12906–12915. https://doi.org/10.1523/JNEUROSCI.2091-11.2011

Blank, H., Kiebel, S. J., & von Kriegstein, K. (2015). How the human brain exchanges information across sensory modalities to recognize other people. *Human Brain Mapping*, *36*, 324–339.

Borowiak, K., Maguinness, C., & von Kriegstein, K. (2019). Dorsal-movement and ventral-form regions are functionally connected during visual-speech recognition. *Human Brain Mapping.*, *41*, 952–972. https://doi.org/10.1002/hbm.24852

Bülthoff, I., & Newell, F. N. (2015). Distinctive voices enhance the visual recognition of unfamiliar faces. *Cognition*, *137*, 9–21. https://doi.org/10.1016/j.cognition.2014.12.006

Bülthoff, I., & Newell, F. N. (2017). Crossmodal priming of unfamiliar faces supports early interactions between voices and faces in person perception. *Visual Cognition*, *25*(4–6), 611–628. https://doi.org/10.1080/02699931.2011.628301

Butcher, N., & Lander, K. (2017). Exploring the motion advantage: Evaluating the contribution of familiarity and differences in facial motion. *Quarterly Journal of Experimental Psychology*, *70*(5), 919–929. https://doi.org/10.1080/17470218.2016.1138974

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

De Renzi, E. (1986). Prosopagnosia in two patients with CT scan evidence of damage confined to the right hemisphere. *Neuropsychologia*, *24*(3), 385–389.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, *137*(3), 1513–1528. https://doi.org/10.1121/1.4906837

Dobs, K., Bülthoff, I., & Schultz, J. (2016). Identity information content depends on the type of facial movement. *Scientific Reports*, *6*, 34301. https://doi.org/10.1038/srep34301

Dobs, K., Ma, W. J., & Reddy, L. (2017). Near-optimal integration of facial form and motion. *Scientific Reports*, *7*, 11002. https://doi.org/10.1038/s41598-017-10885-y

Eger, E., Schyns, P. G., & Kleinschmidt, A. (2004). Scale invariant adaptation in fusiform face-responsive regions. *NeuroImage*, *22*, 232–242. https://doi.org/10.1016/j.neuroimage.2003.12.028

Erber, N. P. (1969). Interaction of audition and vision in the recognition of Oral speech stimuli. *Journal of Speech, Language, and Hearing Research*, *12*(2), 423–425.

Ewbank, M. P., & Andrews, T. J. (2008). Differential sensitivity for viewpoint between familiar and unfamiliar faces in human visual cortex. *NeuroImage*, *40*(4), 1857–1870.

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*(6), 839–849.

Föcker, J., Hölig, C., Best, A., & Röder, B. (2011). Crossmodal interaction of facial and vocal person identity information: An event-related potential study. *Brain Research*, *1385*, 229–245. https://doi.org/10.1016/j.brainres.2011.02.021

Fox, C. J., Iaria, G., & Barton, J. J. S. (2009). Defining the face processing network: Optimization of the functional localizer in fMRI. *Human Brain Mapping*, *30*(5), 1637–1651. https://doi.org/10.1002/hbm.20630

Friston, K. J., Ashburner, J., Kiebel, S., Nichols, T., & Penny, W. (Eds.). (2007). *Statistical parametric mapping: the analysis of functional brain images*. London: Academic Press.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, *6*(3), 218–229. https://doi.org/10.1006/nimg.1997.0291

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, *2*(1–2), 56–78. https://doi.org/10.1002/hbm.460020107

Ghazanfar, A. A., Turesson, H. K., Maier, J. X., van Dinther, R., Patterson, R. D., & Logothetis, N. K. (2007). Vocal-tract resonances as indexical cues in rhesus monkeys. *Current Biology*, *17*(5), 425–430. https://doi.org/10.1016/j.cub.2007.01.029

Girges, C., O'Brien, J., & Spencer, J. (2016). Neural correlates of facial motion perception. *Social Neuroscience*, *11*(3), 311–316.

Girges, C., Spencer, J., & O'Brien, J. (2015). Categorizing identity from facial motion. *The Quarterly Journal of Experimental Psychology*, *68*(April 2015), 1832–1843. https://doi.org/10.1080/17470218.2014.993664

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience*, *7*(5), 555–562. https://doi.org/10.1038/nn1224

He, L., & Dellwo, V. (2016). The role of syllable intensity in between-speaker rhythmic variability. *International Journal of Speech, Language and the Law*, *23*(2), 243–273. https://doi.org/10.1558/ijsll.v23i2.30345

Hölig, C., Föcker, J., Best, A., Röder, B., & Büchel, C. (2014a). Brain systems mediating voice identity processing in blind humans. *Human Brain Mapping*, *35*(9), 4607–4619. https://doi.org/10.1002/hbm.22498

Hölig, C., Föcker, J., Best, A., Röder, B., & Büchel, C. (2014b). Crossmodal plasticity in the fusiform gyrus of late blind individuals during voice recognition. *NeuroImage*, *103*, 374–382. https://doi.org/10.1016/j.neuroimage.2014.09.050

Holm, S. (1979). A simple sequentially Rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Ingram, J. C. L., Prandolini, R., & Ong, S. (1996). Formant trajectories as indices of phonetic variation for speaker identification. *International Journal of Speech Language and the Law*, *3*(1), 129–145. https://doi.org/10.1558/ijsll.v3i1.129

Ives, D. T., Smith, D. R., & Patterson, R. D. (2005). Discrimination of speaker size from syllable phrases. *The Journal of the Acoustical Society of America*, *118*(6), 3816–3822 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16419826

Jenkins, R. E., Tsermentseli, S., Monks, C., Robertson, D. J., Stevenage, S. V, Symons, A. E., & Davis, J. P. (2020). Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. *PsyArXiv*. https://doi.org/10.31234/osf.io/7xdp3

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). "Putting the face to the voice": Matching identity across modality. *Current Biology*, *13*(19), 1709–1714.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specalized for face perception. *The Journal of Neuroscience*, *17*(11), 4302–4311.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B*, *361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2009). Perception and hierarchical dynamics. *Frontiers in Neuroinformatics*, *3*, 20 Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19649171

Kim, C., Shin, H. V., Oh, T. H., Kaspar, A., Elgharib, M., & Matusik, W. (2019). On learning associations of faces and voices. In C. V. Jawahar, H. Li, G. Mori, & K. Schindler (Eds.), *Computer vision – ACCV 2018. ACCV 2018. Lecture notes in computer science* (Vol. 11365, pp. 276–292). Cham: Springer. https://doi.org/10.1007/978-3-030-20873-8_18

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*(18), 1921–1936. https://doi.org/10.1016/S0042-6989(03)00236-0

Knight, B., & Johnston, A. (1997). The role of movement in face recognition. *Visual Cognition*, *4*(3), 265–273. https://doi.org/10.1080/713756764

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., & Pisoni, D. B. (2004). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *30*(2), 378–396. https://doi.org/10.1037/0096-1523.30.2.378.Cross-Modal

Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology*, *12*(4), 259–272.

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition*, *27*(6), 974–985. https://doi.org/10.3758/BF03201228

Lander, K., & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition*, *12*(3), 429–442. https://doi.org/10.1080/13506280444000382

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, *21*(4), R143–R145. https://doi.org/10.1016/j.cub.2010.12.033

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin and Review*, *26*(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, *4*(1), 63–74. https://doi.org/10.1023/A:1009656816383

Leemann, A., Kolly, M. J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, *238*, 59–67. https://doi.org/10.1016/j.forsciint.2014.02.019

Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of face parts and face configurations: An FMRI study. *Journal of Cognitive Neuroscience*, *22*(1), 203–211.

Liu, R. R., Corrow, S. L., Pancaroglu, R., Duchaine, B., & Barton, J. (2015). The processing of voice identity in developmental prosopagnosia. *Cortex*, *71*, 390–397. https://doi.org/10.1016/j.cortex.2015.07.030

Longmore, C. A., & Tree, J. J. (2013). Motion as a cue to face recognition: Evidence from congenital prosopagnosia. *Neuropsychologia*, *51*(5), 864–875. https://doi.org/10.1016/j.neuropsychologia.2013.01.022

Luzzi, S., Coccia, M., Polonara, G., Reverberi, C., Ceravolo, G., Silvestrini, M., … Gainotti, G. (2018). Selective associative phonagnosia after right anterior temporal stroke. *Neuropsychologia*, *31*(116), 154–161. https://doi.org/10.1016/j.neuropsychologia.2017.05.016

Maguinness, C., & Newell, F. N. (2015). Non-rigid, but not rigid, motion interferes with the processing of structural face information in developmental prosopagnosia. *Neuropsychologia*, *70*, 281–295. https://doi.org/10.1016/j.neuropsychologia.2015.02.038

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179–193. https://doi.org/10.1016/j.neuropsychologia.2018.03.039

Maguinness, C., Schall, S., & von Kriegstein, K. (2021). Prior audio-visual learning facilitates auditory-only speech and voice-identity recognition in noisy listening conditions. *PsyArXiv*. https://doi.org/10.31234/osf.io/gc4xa

Maguinness, C., & von Kriegstein, K. (2017). Cross-modal processing of voices and faces in developmental prosopagnosia and developmental phonagnosia. *Visual Cognition*, *25*(4–6), 644–657. https://doi.org/10.1080/13506285.2017.1313347

Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 307–312. https://doi.org/10.1037/a0030945

Mc Dougall, K. (2004). Speaker-specific formany dynamics: An experiment on Australian English /ai/. *Speech, Language and the Law*, *11*(1), 103–130. https://doi.org/10.1097/PSY.0000000000000183

Mc Dougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *Speech, Language and the Law*, *13*(1), 89–126. https://doi.org/10.1097/PSY.0000000000000183

Mc Dougall, K., & Nolan, F. (2007). *Discrimination of Speakers Using the Formant Dynamics of /u:/ in British English*. Proceedings of the 16th International Congress of Phonetic Sciences, pp. 1825–1828. Retrieved from http://www.icphs2007.de

McConachie, H. R. (1976). Developmental prosopagnosia. A single case report. *Cortex*, *12*(1), 76–82. https://doi.org/10.1016/S0010-9452(76)80033-0

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, *56*, 640–662.

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, *12*, 419–446.

Oh, T. H., Dekel, T., Kim, C., Mosseri, I., Freeman, W. T., Rubinstein, M., & Matusik, W. (2019). *Speech2Face: Learning the Face Behind a Voice*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June, pp. 7531–7540. IEEE Computer Society. https://doi.org/10.1109/CVPR.2019.00772

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsycholgia*, *9*, 97–113.

O'Reilly, J. X., Woolrich, M. W., Behrens, T. E., Smith, S. M., & Johansen-Berg, H. (2012). Tools of the trade: Psychophysiological interactions and functional connectivity. *Social Cognitive and Affective Neuroscience*, *7*(5), 604–609.

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences*, *6*(6), 261–266. https://doi.org/10.1016/S1364-6613(02)01908-3

Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, *3*(February), 23. https://doi.org/10.3389/fpsyg.2012.00023

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage*, *56*(4), 2356–2363. https://doi.org/10.1016/j.neuroimage.2011.03.067

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 651–666.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Riedel, P., Ragert, P., Schelinski, S., Kiebel, S. J., & von Kriegstein, K. (2015). Visual face-movement sensitive cortex is relevant for auditory-only speech recognition. *Cortex*, *68*, 86–99. https://doi.org/10.1016/j.cortex.2014.11.016

Roark, D. A., Barrett, S. E., Spence, M., Abdi, H., & O'Toole, A. J. (2003). Memory for moving faces: Psychological and neural perspectives on the role of motion in face recognition. *Behavioral and Cognitive Neuroscience Reviews*, *2*(1), 15–46.

Rosenblum, L. D., Johnson, J. A., & Saldana, H. M. (1996). Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research*, *39*(6), 1159–1170.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Rossion, B. (2014). Understanding face perception by means of prosopagnosia and neuroimaging. *Frontiers in Bioscience*, *6*, 258–307.

Rotshtein, P., Henson, R. N. A., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, *8*(1), 107–113. https://doi.org/10.1038/nn1370

Schall, S., Kiebel, S., Maess, B., & von Kriegstein, K. (2013). Early auditory sensory processing of voices is facilitated by visual mechanisms. *NeuroImage*, *77*, 237–245. https://doi.org/10.1016/j.neuroimage.2013.03.043

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2014). Voice identity recognition: Functional division of the right STS and its behavioural relevance. *Journal of Cognitive Neuroscience*, *27*(2), 280–291.

Schall, S., & von Kriegstein, K. (2014). Functional connectivity between face-movement and speech-intelligibility areas during auditory-only speech perception. *PLoS One*, *9*(1), e86325. https://doi.org/10.1371/journal.pone.0086325

Schelinski, S., Borowiak, K., & von Kriegstein, K. (2016). Temporal voice areas exist in autism spectrum disorder but are dysfunctional for voice identity recognition. *Social Cognitive and Affective Neuroscience*, *11*(11), 1812–1822. https://doi.org/10.1093/scan/nsw089

Schelinski, S., Riedel, P., & von Kriegstein, K. (2014). Visual abiltites are important for auditory-only speech recognition: Evidence from autism spectrum disorder. *Neuropsychologia*, *65*, 1–11.

Schiltz, C., Dricot, L., Goebel, R., & Rossion, B. (2010). Holistic perception of individual faces in the right middle fusiform gyrus as evidenced by the composite face illusion. *Journal of Vision*, *10*(2), 1–16. https://doi.org/10.1167/10.2.25

Schultz, J., & Pilz, K. S. (2009). Natural facial motion enhances cortical responses to faces. *Experimental Brain Research*, *194*(3), 465–475. https://doi.org/10.1007/s00221-009-1721-9

Schweinberger, S. R. (2001). Human brain potential correlates of voice priming and voice recognition. *Neuropsychologia*, *39*(9), 921–936 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11516445

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. https://doi.org/10.1016/j.cortex.2010.11.011

Schweinberger, S. R., & Robertson, D. M. C. (2017). Audiovisual integration in familiar person recognition. *Visual Cognition*, *25*(4–6), 589–610. https://doi.org/10.1080/13506285.2016.1276110

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Sheffert, S. M., & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics*, *66*(2), 352–362 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15129754

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology. Human Perception and Performance*, *28*(6), 1447–1469. https://doi.org/10.1037/0096-1523.28.6.1447

Sidtis, D., & Kreiman, J. (2012). In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication. *Integrative Psychological & Behavioral Science*, *46*(2), 146–159.

Simmons, D., Dorsi, J., Dias, J. W., & Rosenblum, L. D. (2021). Cross-modal transfer of talker-identity learning. *Attention, Perception & Psychophysics*. *83*, 415–434. https://doi.org/10.3758/s13414-020-02141-9

Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex and age. *Journal of the Acoustical Society of America*, *118*, 3177–3186.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America*, *117*(1), 305–318.

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016a). Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, *14*(1), 147470491663031. https://doi.org/10.1177/1474704916630317

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016b). Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, *78*(3), 868–879. https://doi.org/10.3758/s13414-015-1045-8

Steede, L. L., Tree, J. J., & Hole, G. J. (2007). I can't recognize your face but I can recognize its movement. *Cognitive Neuropsychology*, *24*(4), 451–466. https://doi.org/10.1080/02643290701381879

Stevenage, S. V. (2018). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, *116*, 162–178. https://doi.org/10.1016/j.neuropsychologia.2017.07.005

Sumby, W. H., & Pollack, I. (1954). Visual contribution of speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215.

Traunmüller, H., & Eriksson, A. (1994). *The frequency range of the voice fundamental in the speech of male and female adults*. Department of Linguistics, University of Stockholm, Vol. 97, pp. 1905191–1905195.

Tsantani, M., Kriegeskorte, N., McGettigan, C., & Garrido, L. (2019). Faces and voices in the brain: A modality-general person-identity representation in superior temporal sulcus. *NeuroImage*, *201*, 116004. https://doi.org/10.1016/j.neuroimage.2019.07.017

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics*, *13*(1), 19–38.

Van Lancker, D., Kreiman, J., & Wickens, T.-D. (1985). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics*, *13*(1), 39–52.

Voiers, W. D. (1964). Perceptual bases of speaker identity. *The Journal of the Acoustical Society of America*, *36*(6), 1065–1073.

von Kriegstein, K. (2012). A multisensory perspective on human communication. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes*. Boca Raton (FL): CRC Press/Taylor & Francis.

von Kriegstein, K., Dogan, O., Grüter, M., Giraud, A.-L., Kell, C. A., Grüter, T., … Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences*, *105*(18), 6747–6752. https://doi.org/10.1073/pnas.0710826105

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*(1), 48–55.

von Kriegstein, K., & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*(2), 948–955. https://doi.org/10.1016/j.neuroimage.2004.02.020

von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), e326. https://doi.org/10.1371/journal.pbio.0040326 von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*(3), 367–376. https://doi.org/10.1162/0898929053279577

von Kriegstein, K., Kleinschmidt, A., & Giraud, A.-L. (2006). Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex*, *16*(9), 1314–1322. https://doi.org/10.1093/cercor/bhj073

Watson, R., Latinus, M., Charest, I., Crabbe, F., & Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex*, *50*, 125–136. https://doi.org/10.1016/j.cortex.2013.07.011

Weibert, K., & Andrews, T. J. (2015). Activity in the right fusiform face area predicts the behavioural advantage for the perception of familiar faces. *Neuropsychologia*, *75*, 588–596. https://doi.org/10.1016/j.neuropsychologia.2015.07.015

Xu, X., Yue, X., Lescroart, M. D., Biederman, I., & Kim, J. G. (2009). Adaptation in the fusiform face area (FFA): Image or person? *Vision Research*, *49*(23), 2800–2807. https://doi.org/10.1016/j.visres.2009.08.021

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in Cognitive Sciences*, *24*(5), 398–410. https://doi.org/10.1016/j.tics.2020.02.001

Yovel, G., & O'Toole, A. J. (2016). Recognizing people in motion. *Trends in Cognitive Sciences*, *20*(5), 383–395. https://doi.org/10.1016/j.tics.2016.02.005

Zäske, R., Mühl, C., & Schweinberger, S. R. (2015). Benefits for voice learning caused by concurrent faces develop over time. *PLoS One*, *10*(11), 1–12. https://doi.org/10.1371/journal.pone.0143151

Zuo, D., & Mok, P. P. K. (2015). Formant dynamics of bilingual identical twins. *Journal of Phonetics*, *52*, 1–12. https://doi.org/10.1016/j.wocn.2015.03.003

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Maguinness, C., & von Kriegstein, K. (2021). Visual mechanisms for voice-identity recognition flexibly adjust to auditory noise level. *Human Brain Mapping*, *42*(12), 3963–3982. https://doi.org/10.1002/hbm.25532