Language and Cognition (2024), 1–24
doi:10.1017/langcog.2024.3

ARTICLE

Facial cues to anger affect meaning interpretation of subsequent spoken prosody

Caterina Petrone[1], Francesca Carbone[1,2], Nicolas Audibert[3] and Maud Champagne-Lavau[1]

[1]CNRS, LPL, UMR 7309, Aix-Marseille Université, Aix-en-Provence, France
[2]School of Psychology, University of Kent, Canterbury, UK
[3]Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle, Paris, France
Corresponding author: Nicolas Audibert; Email: nicolas.audibert@sorbonne-nouvelle.fr

(Received 15 November 2022; Revised 22 December 2023; Accepted 04 January 2024)

Abstract

In everyday life, visual information often precedes the auditory one, hence influencing its evaluation (e.g., seeing somebody's angry face makes us expect them to speak to us angrily). By using the cross-modal affective paradigm, we investigated the influence of facial gestures when the subsequent acoustic signal is emotionally unclear (neutral or produced with a limited repertoire of cues to anger). Auditory stimuli spoken with angry or neutral prosody were presented in isolation or preceded by pictures showing emotionally related or unrelated facial gestures (angry or neutral faces). In two experiments, participants rated the valence and emotional intensity of the auditory stimuli only. These stimuli were created from acted speech from movies and delexicalized via speech synthesis, then manipulated by partially preserving or degrading their global spectral characteristics. All participants relied on facial cues when the auditory stimuli were acoustically impoverished; however, only a subgroup of participants used angry faces to interpret subsequent neutral prosody. Thus, listeners are sensitive to facial cues for evaluating what they are about to hear, especially when the auditory input is less reliable. These results extend findings on face perception to the auditory domain and confirm inter-individual variability in considering different sources of emotional information.

Keywords: cross-modal affective priming; emotional meaning; facial gestures; French; spoken prosody

1. Introduction

Human emotions are communicated in unimodal or multimodal ways, for example, through the auditory modality (e.g., raising or lowering the pitch of an utterance, e.g., Scherer, 2003), the visual modality (e.g., smiling or frowning, e.g., Ekman, 1992) or both (e.g., Jessen & Kotz, 2013, 2015). In everyday life, visual information often precedes auditory information (Jessen & Kotz, 2013), thus facilitating language processing and comprehension. For instance, simply seeing somebody's angry facial

expression (e.g., via typical facial gestures like lowering and drawing together the eyebrows) makes us expect them to speak to us angrily (Jessen & Kotz, 2013).

In the literature on multimodal emotion perception, the cross-modal affective priming paradigm has often been adopted to investigate to what extent facial gestures impact the subsequent perception of clear vocal expressions of emotions (in the sense of spoken prosodic patterns that unambiguously convey an emotional meaning via their acoustic cues). Here, we use this paradigm to explore how the visual modality (i.e., facial gestures associated with basic emotion expressions) influences the interpretation of subsequent spoken prosody, which varies in the amount of acoustic information conveying affective (angry) meaning. In particular, we target a specific effect already found in the literature on the perception of emotionally degraded or neutral faces (the 'Kuleshov' effect, e.g., Calbi et al., 2017, 2019; Mobbs et al., 2006) to test its potential existence with (emotionally degraded or neutral) audio cues.

1.1. Multimodal perception of emotional prosody

Spoken prosody provides a powerful means to infer other people's emotions (e.g., Banse & Scherer, 1996) that vary in emotional intensity (i.e., strong vs. weak; henceforth 'intensity') and emotional valence (i.e., positive or negative). Multiple phonetic parameters are used to convey basic emotions (e.g., anger, happiness or sadness), such as fundamental frequency (f0), acoustic intensity, speech rate or voice quality-related characteristics like the harmonic-to-noise ratio, jitter and shimmer (e.g., Banse & Scherer, 1996; Gobl & Ní Chasaide, 2003; Scherer, 2003). For instance, (hot) anger is characterized in many languages by an increase in the f0 range, higher intensity, and a faster speech rate or harsh/tense voice compared to a neutral emotional state (e.g., Gobl & Ní Chasaide, 2003; Scherer, 2003). Language-specific phonological features (e.g., choice of specific pitch accents or changes in prosodic phrasing) may also play a role in conveying basic emotions (Cao et al., 2014).

In everyday life, listeners are exposed to multiple sources of emotional information (e.g., auditory and gestural/facial cues), which they may take into account when perceiving emotions (Paulmann & Pell, 2010). Multimodal emotion perception is more efficient than unimodal perception, as evidenced by faster reaction times, higher accuracy or higher emotional ratings (e.g., de Gelder et al., 1999; Pell, 2005; Zhang et al., 2018). A critical factor when investigating multimodal emotion perception is the timing between the sensory stimuli, for example, whether emotional information from the auditory and the visual domain are temporally aligned or whether one stimulus precedes the other (Garrido-Vásquez et al., 2018; Jessen & Kotz, 2013, 2015).

Most of the previous literature on multimodal emotion perception has focused on multimodal integration. In such studies, information from the auditory and visual modalities are presented simultaneously, with the aim of investigating how listeners integrate different sensory inputs into a coherent percept (e.g., de Gelder et al., 1999, 2006; de Gelder & Vroomen, 2000; Massaro & Egan, 1996). In particular, studies on audiovisual integration of prosody have shed considerable light on how auditory and visual information are used *together* for decisions concerning linguistic and paralinguistic contrasts (e.g., Borràs-Comes & Prieto, 2011; Bujok et al., 2022; Crespo Sendra et al., 2013; de Gelder & Vroomen, 2000; House et al., 2001; Massaro & Beskow, 2002; Srinivasan & Massaro, 2003). As for emotion perception, de Gelder and Vroomen

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

(2000) presented an utterance with neutral content which was rendered emotionally ambiguous by manipulating the duration, f0 range and f0 register of the whole utterance between happy and fearful. While the utterance was playing, a picture of a happy or fearful face was simultaneously displayed on the screen. Fearful faces biased the interpretation of the auditory stimuli more toward 'fear', with a stronger effect when the auditory stimuli were ambiguous between the two emotional categories. As de Gelder and Vroomen (2000) argued, the simultaneous presentation of faces and voices pushed listeners to integrate the sensory inputs into a new 'gestalt', with effects mirroring linguistic phenomena of audiovisual integration like the McGurk effect (McGurk & MacDonald, 1976).

Some studies have also pointed to the existence of cross-modal affective priming effects for asynchronous sensory inputs, by which information from one sensory modality influences the processing and interpretation of a signal in another modality that comes into play later (e.g., Garrido-Vásquez et al., 2018; Jessen & Kotz, 2013; Paulmann et al., 2012; Paulmann & Pell, 2010; Pell, 2005). Jessen and Kotz (2013) claimed that affective priming is pervasive during multimodal emotion perception, where visual information from the face often precedes the auditory one. Such studies often adopted the cross-modal affective priming paradigm, by which a prime stimulus from one sensory modality is presented for a specified duration, followed by a target stimulus from another sensory modality. Hence, different from multimodal integration studies, visual and auditory information is presented consecutively rather than simultaneously, with the aim of evaluating the effect of one modality over another one. Thus, priming effects have been explained by psycholinguistic mechanisms of speech perception that are specific to asynchronous stimuli only, like spreading activation from the prime to the target or prime-target congruency check mechanisms (Pell, 2005). Pell (2005) used the cross-modal affective paradigm to study the impact of the preceding spoken prosody on subsequently presented faces. The spoken prosodic patterns and the faces were either combined in congruent pairs (e.g., happy spoken prosody followed by a happy face) or incongruent pairs (e.g., happy spoken prosody followed by an angry face). Emotion recognition was more accurate for congruent than for incongruent prime-target pairs. In the literature, there is also evidence that the processing of spoken emotional prosody is influenced by the preceding emotional face. Studies combining the cross-modal affective paradigm and ERP measurements have found that emotional congruency effects between vocal expressions and facial gestures can be observed very early in time, that is, within the first 250 ms after the onset of the auditory stimulus (e.g., Garrido-Vásquez et al., 2018; Jessen & Kotz, 2013; Paulmann et al., 2009; Pourtois et al., 2002, *inter alia*). Garrido-Vásquez et al. (2018) presented angry, happy or neutral faces followed by pseudo-sentences spoken with angry or happy prosody. They found that prime-target congruency affected the N100 and P200 components, indicating faster and more efficient processing of the spoken emotional prosody. Neurosciences literature has interpreted such effects have pointed to the existence of a process of multimodal emotion perception, which is very different from multimodal integration: cross-modal prediction (Jessen & Kotz, 2013). Specifically, facial gestures would help listeners generating predictions about certain characteristics of a subsequent sound (e.g., in terms of its temporal predictability and content), hence facilitating auditory information processing (Jessen & Kotz, 2013).

A limitation of the literature employing the cross-modal affective paradigm is that it has mostly focused on the perception of clear vocal expressions of emotion. To our

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

knowledge, no studies have explored the effects of facial gestures when the subsequent vocal expressions of emotion are less clear. This question will be addressed in this article.

1.2. The Kuleshov effect

Researchers on face perception agree that basic emotions are signaled via distinctive facial gestures or 'action units' (Ekman, 1992; Ekman et al., 2002). For instance, (hot) anger is typically signaled by lowering and drawing together the eyebrows and widening the eyes, while happiness is typically signaled by smiling (Ekman et al., 2002). Research has revealed an interplay between the clarity of the facial gestures associated to emotions and the emotional context accompanying such gestures (e.g., an emotional scene or situation). When facial gestures are clear and prototypical of a basic emotion, the emotion is read out from the face with no influence of the context. When facial gestures are unclear, people rely more on contextual information to infer emotional meanings (e.g., Aviezer et al., 2008; Carroll & Russell, 1996; Ekman et al., 1982).

The trade-off relationship between contexts and faces has been corroborated by experimental findings on the influence of emotional scenes in the interpretation of faces which are emotionally unclear (i.e., with facial gestures that are not prototypical of a basic emotion) or neutral (i.e., with no visible facial movements). This line of research is mostly based on the Kuleshov effect, named after an early 20th century Soviet filmmaker (Calbi et al., 2017, 2019; Mobbs et al., 2006; Mullennix et al., 2019, *inter alia*). Lev Kuleshov alternated a close-up of a neutral face with pictures of different emotional scenes (e.g., a dead woman or a little girl playing). The face was perceived as expressing an emotion congruent with the preceding context (sadness and happiness, respectively, cf. Calbi et al., 2017).

A first attempt to replicate the original study (Prince & Hensley, 1992) did not find any evidence of the Kuleshov effect. Experimental evidence for the Kuleshov effect has been found in more recent years, though the effect appears to be modulated by different factors (e.g., Calbi et al., 2017, 2019; Mobbs et al., 2006; Mullennix et al., 2019). Mobbs et al. (2006) showed emotional scenes (primes) followed by faces displaying neutral or unclear emotional expressions (targets), consisting of subtle happy and fearful faces. They found that the emotional scenes modulated the ratings of both the neutral and ambiguous faces, but the effects were stronger for emotionally unclear faces than for neutral faces. The authors interpreted this difference as indicating that the effects of the prime were greater when the facial emotion and the emotional scene were emotionally congruent than when a neutral face was paired with an emotional scene. Mullennix et al. (2019) found that the emotional valence of the face was rated more negatively after a negative than a positive or neutral scene. They also reported strong individual variability in the categorization of the neutral face as 'neutral', varying from 19 to 94% across participants.

In sum, while emotional prosody can be conveyed by cues in the auditory domain, its interpretation can depend on visual information. However, it is not yet clear whether the influence of facial gestures is modulated by the clarity of the subsequent vocal emotional expression. Literature on face perception has found an effect of emotional images on the interpretation of unclear or neutral facial expressions. Valence and intensity ratings of target faces depend on the valence of the preceding emotional scenes, with the effect being more variable across individuals for neutral

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

faces. In the present study, we adopted the cross-priming affective paradigm to elucidate the effect of the emotional face on emotional or neutral spoken prosody.

1.3. Methodological limitations of studies on emotional prosody

Previous experimental studies on emotional prosody present two methodological issues which we aim to overcome in the current study. The first issue is that vocal (and visual) expressions of emotions are often elicited from actors by overtly instructing them ('say this sentence angrily'; e.g., Bänziger et al., 2012; Campbell, 2000; Enos & Hirschberg, 2006; Juslin & Laukka, 2003). However, in theater or movies, emotions naturally arise from the actors' deep understanding of their characters using acting techniques (e.g., based on emotional imagination and recall, Enos & Hirschberg, 2006), and they become more and more realistic through rehearsal. This can result in less prototypical, more 'believable' vocal (and visual) expressions than the ones usually obtained in experimental research (Bänziger et al., 2012).

Furthermore, current speech processing methods for isolating acoustic cues to emotional meaning (such as low-pass filters or random splicing) may lead to unnatural and distorted stimuli (Ramus & Mehler, 1999). An alternative method is to use speech synthesis systems. In particular, the MBROLA software (Dutoit et al., 1996) has been used in speech perception research to both dexicalize and manipulate intonational, rhythmic and segmental characteristics of natural speech (Ramus & Mehler, 1999). However, to our knowledge, MBROLA has never been applied to the delexicalization of emotional prosody. We propose to address these methodological limitations in the present article.

1.4. Research goals and hypotheses

Our study investigated the influence of facial gestures on the perception of spoken angry and neutral prosody. In particular, we tested a specific effect in the auditory domain, that is, the Kuleshov effect. In the literature on face perception, this effect has been tested on emotional or neutral faces (Calbi et al., 2019; Mobbs et al., 2006). Here, we evaluated the Kuleshov effect on the evaluation of both angry and neutral spoken prosody. We focused on anger as it provides a signal of potential danger to listeners, drawing their voluntary and involuntary attention toward a threatening situation (Aue et al., 2011). Two behavioral studies were run, in which we applied the cross-modal affective priming paradigm to study whether the emotional information extracted from the visual modality biases the interpretation of a subsequent auditory signal. In our experiments, pictures of faces were used as primes and emotional pseudo-speech stimuli were used as targets. Previous studies using the cross-modal affective paradigm have focused on prosodic emotional expressions with clear acoustic cues. Here, we tested for the first time whether the influence of the preceding emotional face increases as the subsequent auditory stimulus becomes less clear (with a limited repertoire of cues to anger perception or neutrality, i.e., produced without any emotional connotation, Bänziger et al., 2012). For this purpose, we used delexicalized utterances in which we manipulated the amount of acoustic information signaling anger. Specifically, we either deleted or partially preserved global spectral information contained in the utterances, as this is an important acoustic cue for anger perception (Gobl & Ní Chasaide, 2003).

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

The cross-modal affective priming paradigm was combined with rating tasks, in which the auditory stimuli were evaluated in terms of their valence and emotional intensity. Valence and intensity have been often used in the literature to describe basic emotions, in line with bi-dimensional models (Russell, 2003). Previous priming studies have already employed rating tasks for valence and intensity to investigate the influence of affective primes on the emotional meaning evaluation of subsequent stimuli (e.g., Calbi et al., 2019; Flexas et al., 2013).

Our auditory stimuli were resynthesized utterances based on acted speech extracted from French movies to obtain more naturalistic acted speech (Enos & Hirschberg, 2006). Prosodic parameters were controlled and manipulated via speech synthesis, which has the potential to overcome the limitations of low-pass filtering methods (Ramus & Mehler, 1999), still widely applied to the perceptual evaluation of spontaneous emotional speech.

In Experiment 1, we presented angry and neutral auditory stimuli either alone or after a picture of a congruent facial expression. We hypothesized that listeners would judge auditory stimuli produced with angry prosody as more negative and more intense when preceded by an angry facial expression in comparison to the condition in which the stimuli were presented in isolation. This effect was expected to emerge more strongly when the stimuli were acoustically 'impoverished' so that listeners would more actively exploit facial gesture information for the emotional judgment of the subsequent auditory stimuli (de Gelder et al., 2006).

In Experiment 2, we presented angry and neutral auditory stimuli either after a picture of a congruent facial expression (e.g., an angry spoken prosody preceded by an angry face) or an incongruent facial expression (an angry spoken prosody preceded by a neutral face). Neutral facial expressions are uninformative for emotional meaning (Garrido-Vásquez et al., 2018; Jessen & Kotz, 2013). Hence, we expected listeners to judge angry spoken prosody as more negative and more intense when preceded by an angry face than by a neutral face. Furthermore, research on face perception has found that individuals vary strongly in their sensitivity to context when interpreting neutral faces (Mullennix et al., 2019). We thus explored whether individual differences also exist in the perception of neutral spoken prosody. Specifically, we hypothesized that in our sample, some participants would rely more than others on the facial gesture information when judging subsequent neutral prosody.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Two hundred native French speakers (99 women and 101 men, mean age: 27.99 years old, *SD*: 8.01) participated in the experiment. Each participant filled out an online informed consent form before the experiment. Participants were recruited through Prolific and paid five euros. The study was approved by the ethics review board of Aix-Marseille University (2020-12-03-011).

2.1.2. Materials

Auditory stimuli (targets): The auditory stimuli consisted of 36 resynthesized utterances created from 18 natural utterances which were extracted from French movies by a fine arts student (see Supplementary Appendix I). They were selected from a larger set

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

(73 stimuli) after a series of perceptual validation tasks (see par. 2.2.2). The auditory stimuli were converted into. wav files (sampling rate: 48 kHz). The natural utterances were produced by nine French actors (four women, five men) either with (hot) angry (nine utterances) or neutral (nine utterances) spoken prosody. These utterances were 6.54 syllables long on average (mean utterance duration = 0.97 seconds, SD = 0.34). They could also contain words carrying negative or positive valence (e.g., *Je suis pas malade moi!*, 'I am not crazy!'). The natural utterances were of good recording quality and contained very little to no background noise. None of them contained overlapping speech from other actors.

The auditory stimuli were delexicalized to remove the emotional verbal meaning. We created two versions of each utterance that were either poorer or richer in the quantity of emotional prosodic information. In the first version (morphing− condition), only broad phonotactics, rhythm and intonation contours were preserved from the original sentences. In the second condition (morphing+ condition), the delexicalized stimuli were enriched by partially reconstructing the global spectral characteristics of the original utterances. Delexicalization was obtained through MBROLA (Dutoit et al., 1996; Ramus & Mehler, 1999) which performs a resynthesis through the concatenation of diphones using a database of French diphones. We adopted the 'saltanaj' condition (Ramus & Mehler, 1999) in which each phone in the original utterance is substituted by a phone from the same broad phonologic category (e.g., /s/ for fricatives and /a/ for vowels). For the morphing+ condition, the spectral characteristics of the original stimuli were partially reconstructed using a vocal morphing technique through STRAIGHT (Kawahara, 2006) in Matlab (The MathWorks, Inc., Natick, MA). We used the original utterances down-sampled at 16 kHz as the source and the corresponding stimulus in the morphing− condition as the target. Since the segmental durations of the original utterances were preserved in the morphing− condition, a Short-Term Fourier Transform procedure was used on 25 ms overlapping frames of both source and target signals, and the time-aligned morphing process with STRAIGHT was applied to each pair of frames. The morphing rate was set at 0.5 as a trade-off between a lower rate that would have discarded most voice-quality information and a higher rate that would have rebuilt the segmental information of the original utterances. For both conditions, we reconstructed the local variations of intensity of the original utterances via the Vocal Toolkit in Praat (Corretge, 2021). Compared to other existing delexicalization methods based on information removal such as low-pass filtering or wave modulation (Sonntag & Portele, 1998), or copy synthesis with phone substitution (Ramus & Mehler, 1999) as implemented in the morphing− condition, the morphing+ delexicalization method has the advantage of generating more natural utterances in which part of the spectral information relating to voice quality is preserved (Audibert et al., 2023). An example of the morphing manipulation is provided in Figure 1.

Validation tasks for the auditory stimuli: As mentioned previously, our 36 auditory stimuli (18 stimuli × 2 morphing conditions) were selected after perceptual validation tasks. We first validated the 18 stimuli in the morphing+ condition through an intelligibility task (Validation Task 1) and an identification task (Validation Task 2). Another identification task was subsequently run on the corresponding 18 stimuli in the morphing− condition (Validation Task 3). Participants in one task did not take part in the other two tasks nor in the main experiments.

The intelligibility task (Validation Task 1) was performed to ensure that the addition of spectral information did not lead to word recognition. Forty-seven native

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

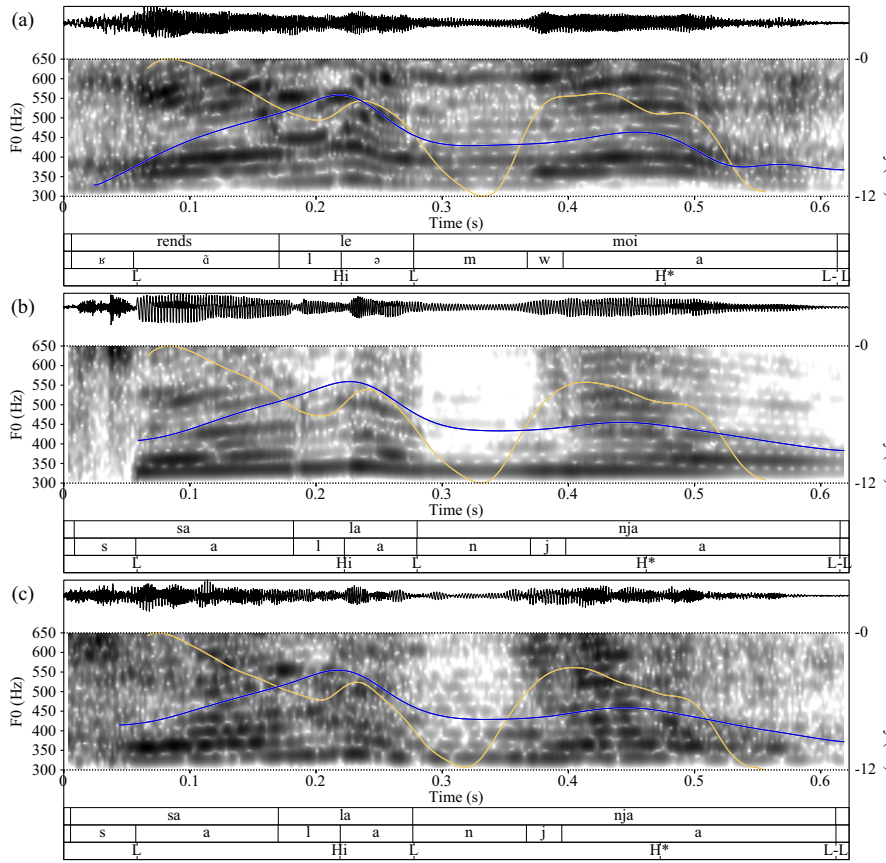

Figure 1. Spectrogram, waveform, smoothed f0 contour (in blue), intensity contour (in yellow) and textgrid for the sentence *Rends-le moi* 'Give it back to me' in the (a) original, (b) morphing− and (c) morphing+ conditions. Textgrids contain the orthographic transcription (tier 1), the IPA transcription (tier 2), and the phonological annotation for the f0 contour within the French ToBI system (tier 3). The dashed lines indicate segmental boundaries. The three conditions match in terms of their intensity contour, f0 contours and phonological annotation (LHi LH* L-L%). The sentence in the example consists of one Intonational Phrase which contains one Accentual Phrase (the basic prosodic unit in French, composed of an early LHi and a late LH* rise, Jun & Fougeron, 2000).

French listeners (38 women and 9 men, mean age: 24.71 years old, *SD*: 7.64), assigned to two counterbalanced lists, transcribed the auditory stimuli orthographically. The stimuli were presented in random order and heard only once. From this task, we retained only those stimuli in the morphing+ condition which had an intelligibility rate equal to or inferior to 30%.

The identification task (Validation Task 2) assessed whether the target emotional states (anger and neutral) were accurately recognized through spoken prosody. Thirty-nine native French listeners (21 women and 18 men, mean age: 30.33 years old, *SD*: 9.59) identified the emotion conveyed through spoken prosody by choosing between 'anger' and 'neutral'. From this task, we retained only those stimuli in the

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

morphing+ condition whose identification rate was equal or superior to 70% (mean identification score: 93%, *SD:* 7%). Based on Validation Tasks 1 and 2, 18 stimuli from the morphing+ condition were selected as our target auditory stimuli as they simultaneously satisfied the selection criteria of both tasks.

We also ran an identification task on the 18 stimuli in the morphing− condition (Validation Task 3). This task assessed whether our spectral manipulation affected emotion recognition. Twenty native French listeners (13 women and 7 men; mean age: 22 years old; *SD*: 8.09) identified the emotion conveyed through spoken prosody by choosing between 'anger' and 'neutral'. Stimuli in the morphing− condition collected lower recognition rates than those in the morphing+ condition (mean identification score: 77%, *SD:* 15%), with the difference being significant [$\beta = 13.97$, SE = 4.35, $t = 3.21$, $p = 0.002$].

In total, 36 utterances were collected [(9 sentences × 2 emotional prosodies (anger, neutral) × 2 morphing conditions (morphing−, morphing+)].

Acoustic analyses: We acoustically analyzed both the original utterances and their resynthesized versions in terms of their mean relative acoustic intensity (dB), utterance duration (seconds) and f0 mean (Hz) using custom scripts in PRAAT (Boersma, 2001). Relative acoustic intensity (dB) was calculated as the difference between the mean acoustic intensity of the penultimate vowel of each utterance, expected to be unaccented, and that of the other vowels within the utterance, to enable comparisons between utterances. The f0 mean was calculated over the entire duration of the utterance. We estimated the overall acoustic difference between the original utterances and their resynthesized versions using Euclidean distances in the 12-dimension space of Mel-frequency cepstral coefficients (MFCC; Davis & Mermelstein, 1980; see also Terasawa et al., 2012 on the relationship between distances in the MFCC space and perceived voice quality). Thirteen coefficients were extracted on 15 ms frames of the speech signal with 5 ms of overlap, before dropping coefficient 0 related to the overall energy in the signal and computing the distance between the original and resynthesized versions on each frame.

Linear regression models showed no differences in acoustic intensity, duration and f0 between the original utterances and the resynthesized stimuli ($p > .05$). The stimuli in the morphing− and morphing+ conditions presented similar values to the original stimuli in terms of their relative acoustic intensity (morphing− vs. original: [$\beta = 0.0031$; $SE = 0.0114$, $t = 0.277$, $p = 0.78$]; morphing+ vs. original: [$\beta = 0.015$; $SE = 0.011$, $t = 1.37$, $p = 0.17$]), utterance duration (morphing− vs. original: [$\beta = -0.0001$; $SE = 0.0116$, $t = -0.013$, $p = 0.99$]; morphing+ vs. original: [$\beta = 0.007$; $SE = 0.011$, $t = 0.62$, $p = 0.53$]) and f0 mean (morphing− vs. original: [$\beta = 1.71$; $SE = 16.3$, t = 0.10, $p = 0.91$]; morphing + vs. original: [$\beta = 0.88$; $SE = 16.3$, $t = 0.05$, $p = 0.95$]). Our spectral manipulation was successful: utterances in the morphing+ condition had lower MFCC coefficients than utterances in morphing− [$\beta = -41.174$; $SE = 3.383$, $t = -12.17$, $p = <0.001$], which confirmed that the stimuli in the morphing+ condition were closer to the originals than those in the morphing− condition in terms of spectral characteristics (Table 1).

Visual stimuli (primes): Eighteen pictures of human faces were selected from the standardized corpus 'Karolinska Directed Emotional Faces database' (KDEF, Lundqvist et al., 1998). The pictures displayed a frontal view of the facial expressions of nine White amateur actors (four women and five men) expressing either (hot) anger or a neutral expression via typical facial gestures. We selected facial expressions presenting high emotion recognition rates (mean for angry faces: 98.21%, *SD*: 4.48; mean for neutral faces: 91.32%, *SD*: 9.72) from previous validation tasks (Goeleven et al., 2008).

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

Table 1. Means and standard deviations (in parentheses) of prosodic parameters for utterances from the original set and for stimuli in the morphing+ and morphing− conditions

Stimuli	Relative acoustic intensity (dB)	Duration (sec.)	f0 mean (Hz)	MFCC distance to original (cepstral magnitude)
Original	0.43 (0.51)	0.96 (0.63)	265.67 (107.12)	N/A
Morphing+	0.47 (0.52)	0.96 (0.63)	264.74 (106.63)	474.83 (115.18)
Morphing−	0.46 (0.51)	0.96 (0.63)	266.75 (107.34)	516.01(148.30)

2.1.3. Procedure

We ran an online experiment via Qualtrics (Snow & Mann, 2013), in which the auditory stimuli were either presented in isolation or after displaying a picture of a face. When the face was present, the visual and the auditory modalities were always paired in a congruent manner (an angry face followed by an angry spoken prosody or a neutral face followed by a neutral spoken prosody). Faces were matched with auditory stimuli according to the sex of the actors (a female face followed by a female voice or a male face followed by a male voice). When the face was absent, the auditory stimuli were preceded by a white background slide. Four counterbalanced lists were constructed which were balanced for spoken prosody (angry and neutral), morphing condition (morphing+ and morphing−), presence of a facial picture (absence and presence) and sex of the actors (female and male). Participants were spread across the four lists so that each auditory stimulus was presented only once. The presentation order was random.

As shown in Figure 2, each trial started with a black fixation cross (500 ms) and was immediately followed by the picture of a face (when the face was present) or by a white screen (when the face was absent) for 1500 ms. When the face was present, it disappeared from the screen before the sound was automatically played. Faces were always in a frontal (straight) view, and this was kept constant across the different pictures. The pictures were presented full screen in the center of the screen and their size was automatically adjusted in Qualtrics in proportion to the screen. The size of the pictures was constant across items. The center of each picture coincided spatially with the starting fixation point. When the picture of the face was absent, the white screen remained while the sound was playing. To limit variations in the visual angle, participants were instructed to sit in front of the computer, to maintain a constant position and to look straight ahead at the fixation point.

Participants evaluated the emotion expressed by the 'speaker's tone of voice' on the valence and intensity scales so that they focused their attention only on the auditory stimuli. Each auditory stimulus was rated on a continuous scale ranging from 'not at all negative' to 'absolutely negative' (for valence) and from 'very weak' to 'very strong' (for emotional intensity). There were no numbers displayed on the slides, but responses were recorded as scores ranging from 1 to 100 (Bhatara et al., 2016), with 1 indicating 'not at all negative' (for valence) and 'very weak' (for intensity) and 100 indicating 'absolutely negative' (for valence) and 'very strong" (for intensity).

2.1.4. Statistical analysis

All statistical tests were performed in R version 4.1.1 (R Core Team, 2022). The package *lmerTest* version 3.1.3 (Kuznetsova et al., 2017) was used for running mixed

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

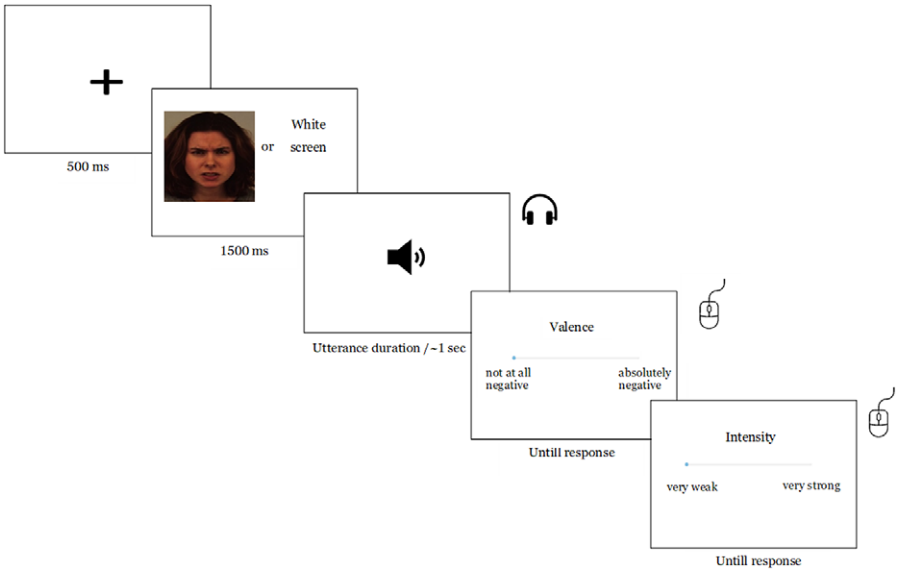

Figure 2. Illustration of the experimental paradigm.

models, while *ggplot2* version 3.3.6 (Wickham, 2016) was used for graphical exploration. Linear mixed effects models were run to analyze valence and intensity scores as a function of SPOKEN PROSODY (neutral vs. angry), FACE (presence vs. absence) and MORPHING (morphing+ vs. morphing−). In addition, PARTICIPANT SEX (female vs. male) and ACTOR SEX (female vs. male) were included as control variables. Interactions among all five fixed factors were included as well. PARTICIPANT (1–200) and ACTOR (1–9) were the random intercepts. We started the statistical analysis by fitting each model with the two intercepts, and by including by-participant and by-actor random slopes for spoken prosody, face and morphing. Since the models showed convergence issues, we simplified the random structure of the models to reduce overparametrization (e.g., by deleting random components with very little variance). The simplification of the random structure of the model did not change the interpretation of the results. To better understand possible interactions (e.g., effects of MORPHING across SPOKEN PROSODY), we ran the models three times, changing the reference level (intercept) for SPOKEN PROSODY and MORPHING. Thus, the cut-off point for significance was set at 0.016 [$p = 0.05$ divided by the number of models (3) run]. Full model outputs are given in Supplementary Appendix II.

The final model for both valence and intensity scores was:

$$\begin{aligned}\text{Valence}(\text{or Intensity}) \sim \text{Spoken\_Prosody} \times \text{Morphing} \times \text{Face} \times \text{Listener\_sex} \times \\ \text{Actor\_sex} + (1 + \text{Spoken\_Prosody} + \text{Morphing} + \text{Face} \mid \text{Participant}) + \\ (1 + \text{Spoken\_Prosody} \mid \text{Actor}).\end{aligned}$$

2.2. Results

We found an effect of SPOKEN PROSODY, in that auditory stimuli produced with angry spoken prosody were judged as more negative [$\beta = 20.98$, $SE = 2.54$, $t = 8.26$,

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

$p < .001$] and more intense [$\beta = 19.6$, $SE = 2.53$, $t = 7.74$, $p < .001$] than those produced with neutral spoken prosody. Specifically, angry and neutral spoken prosody scored on average 61.7 and 22.4 on the valence scale, respectively; and they scored 61.5 and 25.5 on the intensity scale, respectively (Figure 3). The factor MORPHING further modulated the judgments of angry spoken prosody, but not those of neutral spoken prosody. Angry spoken prosody in the morphing+ condition was judged as more negative [$\beta = 22.36$, $SE = 2.96$, $t = 7.55$, $p < .001$] and more intense [$\beta = 18.29$, $SE = 2.45$, $t = 7.47$, $p < .001$] than angry spoken prosody in the morphing− condition. On average, angry spoken prosody in the morphing+ and in the morphing− conditions scored 73.6 and 49.7 on the valence scale, respectively, and it scored 72.08 and 51.04 on the intensity scale. Conversely, the morphing manipulation had no significant effects on neutral spoken prosody for valence or intensity. Finally, for the valence scale only, the factor FACE modulated the judgment of angry spoken prosody, but this effect was limited to auditory stimuli in the morphing− condition [$\beta = 6.24$, $SE = 2.24$, $t = 2.79$, $p = .005$]. As shown in Figure 3, when an angry face preceded an auditory stimulus with 'impoverished' angry spoken prosody (morphing− condition), the valence score was slightly more negative (mean score = 52.6) compared to the condition in which the same stimulus was presented in isolation (mean score = 46.8). The re-leveled model showed that the effect of FACE on valence scores was not significant in the morphing+ condition ($p = 0.36$). Furthermore, there was no effect of FACE before neutral spoken prosody ($p = 0.41$). Finally, PARTICIPANT SEX and ACTOR SEX were not significant for valence or intensity. Their interactions with SPOKEN PROSODY, MORPHING and FACE

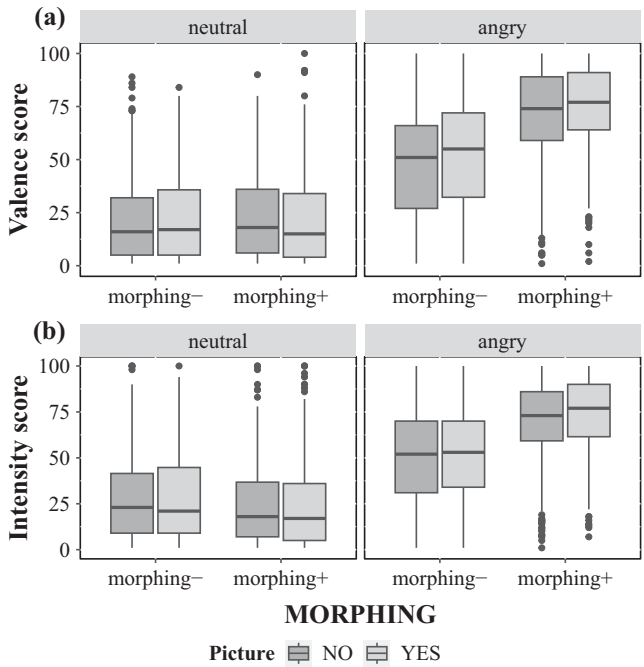

Figure 3. Boxplots of valence (a) and intensity (b) score across MORPHING, split by SPOKEN PROSODY and FACE.

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

were also not significant. The output of the linear mixed models is presented in Supplementary Appendix II.

2.3. Discussion

Experiment 1 showed that ratings for valence and intensity depended on whether the auditory stimuli were produced with angry versus neutral prosody. Judgments of angry spoken prosody were further modulated by the morphing manipulation, that is, whether or not angry auditory stimuli contained spectral information from the original utterances. Crucially, facial expressions affected valence judgments of subsequent angry spoken prosody, but the effect was limited to auditory stimuli containing no spectral information.

The rating task reflects the general assessment of emotions along the valence and intensity dimensions, with angry spoken prosody being judged as more negative and more intense than neutral spoken prosody (Bradley & Lang, 2000). Our auditory stimuli were based on audio excerpts from French movies. This result thus extends findings from acted speech elicited in the lab to acted speech elicited in more ecological settings (Enos & Hirschberg, 2006). Angry spoken prosody was judged as more negative and more intense in the morphing+ than in the morphing− condition, while there was no difference for neutral spoken prosody across the morphing conditions. The crucial difference between these two conditions was the presence versus absence of voice quality-related acoustic features. This result confirms previous research, which found that voice quality is crucial for the perception of anger (Gobl & Ní Chasaide, 2003). Valence and intensity rating scores for angry auditory stimuli in the morphing− condition were in between neutral auditory stimuli and angry auditory stimuli in the morphing+ condition, providing further evidence that f0 and intensity are also reliable cues to anger (Scherer, 2003).

Being presented with a picture of an angry facial expression led to more negative judgments of the valence of 'impoverished' angry auditory stimuli. As stimuli in the morphing− condition conveyed the speaker's emotional state less clearly, the angry facial expression may have resulted in negative visual information affecting the judgment of the following auditory stimulus. Stimuli in the morphing+ condition contained a higher number of acoustic cues for angry spoken prosody. Hence, participants did not need to rely on preceding facial information for judgments of emotional spoken prosody for this subset of stimuli.

The observed effect of FACE was rather small in size. In the standard affective priming paradigm, targets are always preceded by either emotionally congruent or incongruent primes. Here, participants were exposed to a mixed block as the auditory stimuli were either preceded by congruent primes or presented in isolation. We considered the absence of a face as our control condition, and we wanted to compare this condition with a condition in which the auditory stimuli were preceded by emotionally congruent faces. Because the instructions focused participants' attention on the targets, the mixed block may have further attenuated the impact of task-irrelevant distracting information (the facial gestures). This limitation was overcome in Experiment 2, in which primes always preceded the targets and they could create either congruent or incongruent prime-target pairs. This design is more in line with standard affective priming studies (e.g., Paulmann & Pell, 2010).

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

3. Experiment 2

Experiment 2 broadly followed the experimental procedure of Experiment 1, with two major differences. First, in Experiment 2, we used only the auditory stimuli in the morphing− condition as Experiment 1 showed that facial gestures only have an effect in this condition. Second, the auditory stimuli in Experiment 2 were always preceded by a picture of a face, which could be either emotionally congruent or incongruent.

3.1. Methods and materials

3.1.1. Participants

Sixty-seven native French speakers between the ages of 18 and 55 (48 women and 19 men, mean age: 24.74 years old, *SD*: 9.86) who did not take part in previous tasks participated in the experiment. Each participant signed an online informed consent form before the start of the experiment.

3.1.2. Materials and procedure

The target stimuli consisted of the 18 resynthesized utterances from the morphing− condition in Experiment 1. The stimuli expressed either anger (nine stimuli) or neutrality (nine stimuli) through spoken prosody. The primes consisted of the 18 pictures from the KDEF (Lundqvist et al., 1998) employed in Experiment 1, showing facial expressions with typical facial gestures associated to either anger or neutrality. The setting of the experiment and the procedure were identical to Experiment 1. In contrast to Experiment 1, the condition in which auditory stimuli were presented in isolation was excluded. Each auditory stimulus (target) was preceded by an angry or a neutral face (primes), leading to emotional prime-target pairs which were congruent (angry face/angry spoken prosody or neutral face/neutral spoken prosody) or incongruent (angry face/neutral spoken prosody or neutral face/angry spoken prosody). Facial expressions were matched with the auditory stimuli according to the sex of the actors (female faces were paired with female voices and male faces with male voices). Two counterbalanced lists were constructed, each composed of 18 trials balanced for emotion of spoken prosody (angry and neutral), emotion of face (angry and neutral) and sex of the actors (female and male). Participants were instructed to rate the speakers' tone of voice on the valence and intensity scales. They were presented with each visual–auditory stimuli pair only once. Visual–auditory pairs were presented in random order within each list.

3.1.3. Statistical analyses

Statistical tests were performed in R (R Core Team, 2022). To explore whether some listeners relied more on acoustic cues than on the facial expression to judge spoken prosody, we performed a hierarchical cluster analysis based on valence judgments of neutral spoken prosody. We focused on this condition as we expected larger individual differences in the judgment of neutral spoken prosody based on research on the Kuleshov effect (Mullennix et al., 2019). We first computed a distance matrix using the Euclidean distance to quantify the distance among individuals. Specifically, we computed the distance between mean valence judgments of neutral spoken prosody preceded by angry faces and mean valence judgments of neutral spoken prosody preceded by neutral faces. Then, a hierarchical cluster analysis was

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

conducted on this distance matrix via the Ward algorithm (Everitt et al., 2011; Ward, 1963). The Ward algorithm was used as a linkage method, that is, it allowed us to classify our sample of participants into subgroups such that similar participants were grouped in the same subgroup. More precisely, the Ward's method is an agglomerative method, which combines clusters whose grouping leads to the minimum increase in total within cluster variance. This method is standard in clustering analyses to model individual variability (e.g., Rivière et al., 2018) and it is preferred because it minimizes the increase in the total within-cluster sum of squared error. This increase is proportional to the squared Euclidean distance between cluster centers. Graphical exploration of the data structure was made via a dendrogram. The optimal number of clusters was defined using the gap statistic (Tibshirani et al., 2001), that compares the total intracluster variation for different values of k with their expected values under a reference null distribution of the data (i.e., a distribution with no obvious clustering). The reference distribution is generated using Monte Carlo simulations of the sampling process. In our study, the gap statistic was applied to the widely used k-mean clustering (Tibshirani et al., 2001). The gap statistic found two main clusters. The two-cluster subgrouping indicated that our participants had two main patterns of performance (Supplementary Appendix III). The two clusters contained 31 and 36 participants, respectively, and they did not differ in age ($p > 0.05$) or sex ($p > 0.05$). Specifically, the mean age was 23.7 years old (for cluster 1) and 24.5 years old (for cluster 2). There were 21 women, 9 men and 1 non-binary for cluster 1; and 26 women, 9 men and 1 non-binary for cluster 2. Linear mixed models were run to analyze valence and intensity scores as a function of SPOKEN PROSODY (neutral vs. angry), FACE (neutral vs. angry), CLUSTER (cluster 1vs. cluster 2) and their interactions. Experiment 1 revealed no effects on the sex of the participant and the actor. These two control variables were thus not included in the analysis of results for Experiment 2. To better understand possible interactions, we ran the models three times by re-leveling the intercept (alpha level = 0.016). Following simplification procedures like the ones described in Section 2.1.4, the final model was:

$$\text{Valence}\,(\text{or Intensity}) \sim \text{Spoken\_Prosody} \times \text{Face} \times \text{Cluster} + (1\,|\,\text{Participant}) + (1 + \text{Face}\,|\,\text{Actor}).$$

3.2. Results

Based on the hierarchical clustering analysis, we found different patterns of performance depending on listeners' use of preceding facial information (Figure 4). Participants in cluster 1 judged angry spoken prosody as more negative and more intense when preceded by an angry face (valence score: 64.78; intensity score: 62.31) than by a neutral face (valence score: 44.48; intensity score: 52.24). The effect of FACE on angry spoken prosody was significant both for valence [β = 20.67, SE = 2.99, t = 6.91, $p < 0.001$] and intensity [$\beta = 10.59$, $SE = 2.83$, $t = 3.73$, $p < 0.001$]. Furthermore, cluster 1 rated neutral spoken prosody as more negative and more intense when preceded by an angry face (valence score: 54.76; intensity score: 37.41) than by a neutral face (valence score: 29.65; intensity score: 27.78). The re-leveled model showed that the FACE effect on neutral spoken prosody was significant both for valence (β = 25.48, SE = 2.99, t = 8.52, $p < 0.001$) and intensity (β = 10.14, SE = 2.83, t = 3.57, $p < 0.001$). The size of the FACE effect was very similar for both neutral and angry spoken prosody. For valence ratings, the difference between congruent and incongruent visual–auditory pairs was 4.81 points higher for neutral than for angry spoken

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

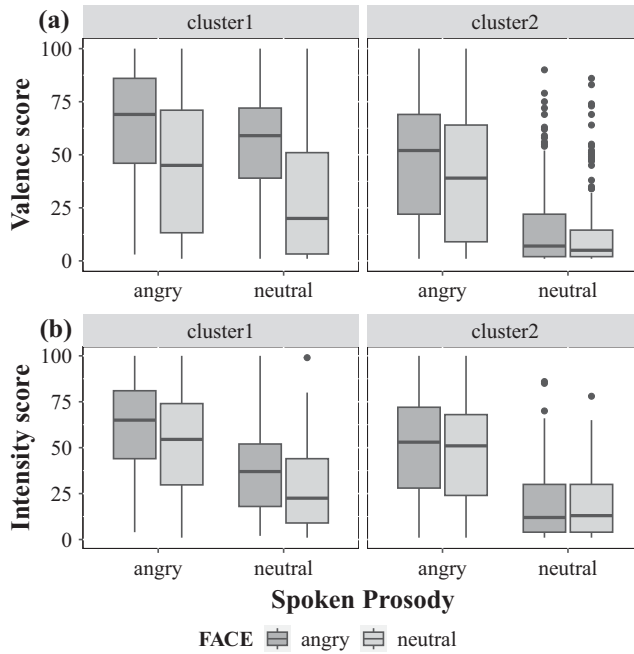

Figure 4. Boxplots of valence (a) and intensity (b) scores across SPOKEN PROSODY, split by CLUSTER and FACE.

prosody ($\beta = 46.16$, $SE = 4.16$, $t = 11.09$, $p < 0.001$). For intensity ratings, the difference between congruent and incongruent visual–auditory pairs was only 0.43 of a point higher for angry spoken prosody than for neutral spoken prosody ($\beta = 20.73$, $SE = 3.5$, $t = -5.90$, $p < 0.001$). Participants in the cluster 2 subgroup were only sensitive to angry faces before angry spoken prosody for valence ratings, with a smaller effect compared to cluster 1 ($\beta = -17.04$, $SE = 3.53$, $t = -4.81$, $p < 0.001$). No effect of FACE on angry spoken prosody was found for intensity ratings ($p = 0.021$). Furthermore, ratings of neutral spoken prosody preceded by an angry face did not differ from ratings of neutral spoken prosody preceded by neutral faces for valence ($p = 0.26$) or intensity ($p = 0.75$). The output of the linear mixed models is presented in Supplementary Appendix IV.

3.3. General discussion

Experiments 1 and 2 showed consistent findings, as the evaluation of (impoverished) angry and neutral spoken prosody was affected by the preceding angry face. Experiment 2 also showed the existence of two clusters of participants, who used facial information either more or less to judge the valence and the intensity of speakers' spoken prosody. Facial gestures associated to anger led participants in cluster 1 to judge the subsequent spoken prosody as more negative and more intense, independently of whether the spoken prosody was angry or neutral. For participants in cluster 2, the angry face led to more negative judgments of angry spoken prosody only. The magnitude of the effect was smaller than that found for cluster 1, and it was limited to the valence scale.

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

The effect of angry facial gestures on angry spoken prosody confirms several previous findings on cross-modal affective priming (see Garrido-Vásquez et al., 2018 and references therein): facial gestures associated to emotional expressions impact the interpretation of the following target more strongly when the prime-target pair is emotionally congruent. Furthermore, in our study, the angry spoken prosody was less salient than the clear angry facial expression preceding it as it was acoustically impoverished. Hence, cues to angry meaning were 'reinforced' by the preceding congruent visual information. This result parallels findings both on multimodal emotion integration (de Gelder & Vroomen, 2000; Massaro & Egan, 1996; Pourtois et al., 2002) and on multimodal integration of intonational pragmatic variables like incredulity or contrastive focus (e.g., Borràs-Comes & Prieto, 2011; Crespo Sendra et al., 2013; Prieto et al., 2015). All these studies suggest that listeners are sensitive to the relative weight of visual and auditory cues, with the effect of one sensory input being stronger when the other sensory input is more ambiguous.

Note though that multimodal integration studies are based on the simultaneous presentation of different sensory inputs, which are thus integrated by listeners into a single percept. The aim of such studies is to evaluate how important cues from one modality are compared to cues from another modality when making, for example, linguistic or paralinguistic decisions (e.g., for prosody: Borràs-Comes & Prieto, 2011; Bujok et al., 2022; Crespo Sendra et al., 2013; de Gelder & Vroomen, 2000; House et al., 2001; Massaro & Beskow, 2002; Srinivasan & Massaro, 2003). Thus, findings on synchronous displays of different sensory inputs have been discussed in connection with integration models for information processing (e.g., the fuzzy logical model, Massaro & Cohen, 1993; the weighted averaging model, Massaro, 1998; the additive model, Cutting et al., 1992).

Our study, on the other side, is based on the asynchronous presentation of different sensory inputs. Thus, we tested priming effects independent of audiovisual integration, as we showed that gestural faces influence the emotional evaluation of spoken prosody even when information from the two modalities is not temporally aligned. For instance, priming effects can result spreading activation or congruency check mechanisms (Pell, 2005). Visual–auditory congruity effects may reflect the fact that information from the visual modality (e.g., the angry face) activates certain emotional features that are shared to some extent by the subsequent emotional expressions of emotion (e.g., the angry spoken prosody), hence facilitating its accessibility. Paulmann and Pell (2010) and Pell (2005) hypothesized that different sensory inputs (such as angry spoken prosody and angry faces) are organized around the same emotion concepts (e.g., 'anger', e.g., Paulmann & Pell, 2010; Pell, 2005) in, for example, a semantic memory network. The presentation of an emotional or affective prime could induce an automatic affective evaluation which would spread from the primes to the emotionally congruent targets within the network. Alternatively, affective priming effects could be explained by listeners' tendencies to judge the compatibility of the affective components of the prime-target pair (e.g., Calbi et al., 2017). Congruency effects could explain why, in Experiment 2, we found more systematic effects for anger-anger target pairs (in which the prime and the target are affectively congruent), while there was much more inter-speaker variability for anger-neutral target pairs (in which the prime and the target are not affectively congruent). Congruency mechanisms have been also invoked in the literature on the Kuleshov effect for interpreting neutral face perception (Calbi et al., 2019).

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

Another important contribution of our study compared to previous ones is related to our focus on spectral auditory cues. These cues have been almost neglected in the literature on multimodal perception and, in particular, in the literature on audio-visual integration employing simultaneously presented sensory inputs. Most of the previous studies have focused on the contribution of auditory cues such as f0, amplitude or duration (e.g., Bujok et al., 2022; de Gelder & Vroomen, 2000; Massaro & Beskow, 2002). However, spectral cues are crucial for auditory prosody perception, and, in particular for anger perception (Gobl & Ní Chasaide, 2003). By combining ecological valid materials (utterances extracted from movies) and sophisticated delexication methods (MBROLA and STRAIGHT), we were able to show that nuances in voice quality-related characteristics can modulate the influence of emotional faces on the evaluation of auditory prosody even when other cues (e.g., f0 and duration) are preserved.

Interestingly, neuroscience studies claimed that cross-modal affective priming effects can be due to cross-modal prediction (Jessen & Kotz, 2013). Specifically, emotionally congruent facial gestures facilitate the processing of clear emotional auditory information, both at the content and at the temporal levels (e.g., Garrido-Vásquez et al., 2018; Jessen & Kotz, 2013, 2015). In turn, cross-modal emotion prediction is considered an instance of predictive coding (Jessen & Kotz, 2013), allowing faster processing of multiple sources of information at both linguistic and extralinguistic levels (Hagoort, 2019; Huettig & Mani, 2016; Paulmann & Pell, 2010; Pickering & Garrod, 2007; Van Berkum et al., 2005; see also Corps, 2018 and references therein). While our study is based on behavioral measures only (rating tasks), it is in line with the idea that emotional facial gestures may predict auditory information at the content level, as the perception of angry faces prime the perception of angry voices (Jessen & Kotz, 2013).

Compared to previous cross-modal affective priming studies (thus, with asynchronous presentation of different sensory inputs), we showed for the first time that the influence of facial gestures on the evaluation of the subsequent acoustic signal emerges more strongly at the behavioral level when vocal expressions of emotions are acoustically less clear. In the future, it will be necessary to run electrophysiological studies to investigate to what extent such priming effects for acoustically impoverished/neutral auditory signals results from cross-modal prediction. This would eventually support the claim that listeners use predictions especially when required to compensate for ambiguous inputs, due to strong top-down influences on interpretation in such cases (Huettig & Mani, 2016; Pickering & Garrod, 2007). Thus, our findings may be also in line with the constraint satisfaction model (Degen & Tanenhaus, 2019). According to this model, we generate expectations about meaning interpretation based on multiple sources of information from visual or auditory modality processed rapidly in a weighted manner, as soon as they are available. Bottom-up information from the linguistic signal is combined with top-down expectations to determine how incoming information is interpreted (Degen & Tanenhaus, 2019).

Experiment 2 showed that contrary to participants in cluster 1, participants in cluster 2 did not use facial gesture information to interpret neutral spoken prosody. These results confirm the existence of inter-individual variability in considering the different sources of information (e.g., visual and acoustic cues), with individuals being more or less sensitive to such cues (Hamann & Canli, 2004; Jun & Bishop, 2015; Rivière et al., 2018). In particular, discrepancies in the evaluation of neutral spoken

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

prosody parallel EEG findings using the cross-modal affective priming paradigm (Garrido-Vásquez et al., 2018; Jessen & Kotz, 2013). Garrido-Vásquez et al. (2018) found that when the visual–auditory pair includes a neutral stimulus, this may either lead to the perception of some kind of audiovisual congruency (with facilitation effects) or audiovisual incongruency (with interference or no facilitation effects). This would explain why cluster 1 showed effects for angry face-neutral spoken prosody pairs, which were comparable to those for angry face-angry spoken prosody pairs (as the pair was interpreted as congruent), while cluster 2 showed no effects for angry face-neutral spoken prosody pairs (the pair was interpreted as incongruent).

Our results on the emotional evaluation of neutral spoken prosody are in with the literature regarding the effects of emotional scenes on neutral expressions ('Kuleshov' effect). Results obtained on (impoverished) angry and neutral spoken prosody extend findings from the visual domain to the auditory one, suggesting the existence of an 'auditory' Kuleshov effect on the perception of spoken prosody. We showed that this effect applies even when the sensory inputs come from different domains (see Baranowski & Hecht, 2017 for similar results on faces and music). The fact that some people use facial cues more than others is also consistent with Kim et al. (2022) and Mullennix et al. (2019), who reported strong individual differences for context-dependency in the perception of neutral faces. Note though that it is difficult to make a straightforward comparison between our results on neutral prosody in the present study and results on neutral faces in the Kuleshov literature, as we lack a condition in which use emotional scenes as primes instead of emotional faces. To our knowledge, there are no studies so far comparing the effects of faces versus scenes. This kind of comparison is relevant to understand whether the effects we found in the current study are specific to faces and voices during multimodal emotion perception, or whether they are interpretable in terms of more general contextual effects on voice processing.

Concerning effect sizes, smaller effects were found in our study compared to multimodal integration studies. There may be at least two different reasons for this. First, the priming paradigm may have attenuated the effects, as we know that behavioral priming effects are usually quite small (Weingarten et al., 2016). Furthermore, in multimodal integration studies, stimuli are often manipulated in a continuum from one category to another across gradual steps of manipulation (e.g., Crespo Sendra et al., 2013; de Gelder & Vroomen, 2000). In our study, on the other hand, 'impoverished' stimuli (i.e., with no spectral cues) contained strong f0, durational and intensity cues to anger, that biased judgments toward a more negative and more arousing evaluation even when presented in isolation (as shown in Experiment 1).

Finally, the effect of the face was either null (Experiment 1 and cluster 2) or smaller (cluster 1) in the intensity than in the valence scale. In our study, the intensity scale was presented after participants responded in their own time to the valence scale (following Calbi et al., 2019). Hence, facial emotional features may have faded from participants' memories within a few seconds, explaining the smaller or null face effects on intensity.

It should be noted that the generalizability of our results is limited by the fact that, because of methodological constraints (we used naturalistic acted speech), we focused only on one emotion. It is necessary to compare the relative weight of different cues on priming effects across several emotions. Listeners variably rely on visual versus vocal cues for emotion recognition, depending on the specific basic emotion (Scherer, 2003). Thus, priming effects may be stronger for emotions that are

https://doi.org/10.1017/langcog.2024.3 Published online by Cambridge University Press

more clearly expressed via visual cues. On the other side, as we already stated, our study was aimed at investigating a specific effect, that is, the Kuleshov effect, in the auditory domain. In the literature on face perception, this effect has been tested either on neutral faces (Calbi et al., 2019) or on emotional faces (Mobbs et al., 2006). In a similar vein, we tested the Kuleshov effect on both neutral and emotional (angry) auditory prosody. In this respect, our main result is that angry faces affect the evaluation of both neutral and impoverished emotional auditory signals. The effect was more robust for anger-anger prime-target pairs while it was more listener-dependent for anger-neutral prime-target pairs.

In conclusion, the present study employed a cross-modal affective priming paradigm with pictures of static facial expressions followed by spoken neutral and angry prosody. Facial gestures to anger influenced the evaluation of (impoverished) angry and neutral spoken prosody suggesting that multimodal emotion perception applies even when sensory inputs are not temporally aligned. We also showed that, while all participants use visual information when paired with congruent emotional spoken prosody, only some of them rely on facial gestures to interpret neutral spoken prosody. We think that cross-modal affective priming may play a facilitatory role in everyday communication, for example, it could enable emotional predictions, based on facial cues, of what we are about to hear. Future studies should clarify which are the neural mechanisms underlying the influence of emotional faces on the processing of neutral and acoustically impoverished emotional prosody, and which is the source of individual variability in using facial- versus signal-based information.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/langcog.2024.3.

Data availability statement. Results of the perception studies and R codes are available on the OSF: https://osf.io/ca9uz/. French legislation does not allow us to deposit our audio stimuli, which are film extracts, on an open-access platform. For more information, please contact the first author.

Funding statement. This study was supported by an A*MIDEX (Aix-Marseille University) grant to C.P. and M.C.-L.

Competing interest. The authors declare none.

References

Audibert, N., Carbone, F., Champagne-Lavau, M., Housseini, A. S., & Petrone, C. (2023). Evaluation of delexicalization methods for research on emotional speech. In *Proceedings of Interspeech* (pp. 2618–2622). HAL Open Science. https://doi.org/10.21437/Interspeech.2023-1903

Aue, T., Cuny, C., Sander, D., & Grandjean, D. (2011). Peripheral responses to attended and unattended angry prosody: A dichotic listening paradigm. *Psychophysiology*, 48, 385–392. https://doi.org/10.1111/j.1469-8986.2010.01064.x

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., & Bentin, S. (2008). Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychological Science*, 19(7), 724–732. https://doi.org/10.1111/j.1467-9280.2008.02148

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3), 614. https://doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception. *Emotion*, 12(5), 1161–1179. https://doi.org/10.1037/a0025827

Baranowski, A. M., & Hecht, H. (2017). The auditory Kuleshov effect: Multisensory integration in movie editing. *Perception*, 46(5), 624–631. https://doi.org/10.1177/0301006616682754

Bhatara, A., Laukka, P., Boll-Avetisyan, N., Granjon, L., Anger Elfenbein, H., & Bänziger, T. (2016). Second language ability and emotional prosody perception. *PLoS One*, 11(6), e0156855. https://doi.org/10.1371/journal.pone.0156855

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9), 341–345.

Borràs-Comes, J., & Prieto, P. (2011). 'Seeing tunes.' The role of visual gestures in tune interpretation. *Laboratory Phonology*, 2(2), 355–380. https://doi.org/10.1515/labphon.2011.013

Bradley, M. M., & Lang, P. J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology*, 37(2), 204–215. https://doi.org/10.1111/1469-8986.3720204

Bujok, R., Meyer, A. S., & Bosker, H. R. (2022). Audiovisual perception of lexical stress: Beat gestures are stronger visual cues for lexical stress than visible articulatory cues on the face. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/y9jck

Calbi, M., Heimann, K., Barratt, D., Siri, F., Umiltà, M. A., & Gallese, V. (2017). How context influences our perception of emotional faces: A behavioral study on the Kuleshov effect. *Frontiers in Psychology*, 8, 1684. https://doi.org/10.3389/fpsyg.2017.01684

Calbi, M., Siri, F., Heimann, K., Barratt, D., Gallese, V., Kolesnikov, A., & Umiltà, M. A. (2019). How context influences the interpretation of facial expressions: A source localization high-density EEG study on the 'Kuleshov effect'. *Science Reports*, 9, 2107. https://doi.org/10.1038/s41598-018-37786-y

Campbell, N. (2000). Databases of emotional speech. In *Proceedings of the ISCA workshop on speech and emotion* (pp. 34–38). ICSA. https://doi.org/10.1016/s0167-6393(02)00070-5

Cao, H., Beňuš, Š., Gur, R. C., Verma R., & Nenkova, A. (2014). Prosodic cues for emotion: Analysis with discrete characterization of intonation. *Proceedings of speech prosody*, 2014, 130–134. https://doi.org/10.21437/SpeechProsody.2014-14

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions express specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 70, 205–218. https://doi.org/10.1037/0022-3514.70.2.205

Corps, R. E. (2018). *Coordinating utterances during conversational dialogue: The role of content and timing predictions.* PhD Thesis, The University of Edinburgh, Edinburgh.

Corretge, R. (2021). Praat vocal toolkit: A Praat plugin with automated scripts for voice processing [software package]. http://www.praatvocaltoolkit.com/index.html

Cutting, J.E., Bruno, N., Brady, N.P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, 121, 362–381.

Crespo Sendra, V., Kaland, C. C. L., Swerts, M. G. J., & Prieto, P. (2013). Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics*, 47(1), 1–13. https://doi.org/10.1016/j.pragma.2012.08.008

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

de Gelder, B., Meeren, H. K., Righart, R., van den Stock, J., van de Riet, W. A., & Tamietto, M. (2006). Beyond the face: Exploring rapid influences of context on face processing. *Progress in Brain Research*, 155, 37–48. https://doi.org/10.1016/S0079-6123(06)55003-4

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, 14, 289–311. https://doi.org/10.1080/026999300378824

de Gelder, B., Vroomen, J., & Pourtois, G. R. C. (1999). Seeing cries and hearing smiles: Cross-modal perception of emotional expressions. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events (Advances in psychology* (Vol. 129, pp. 425–438). Elsevier Science Publishers. https://doi.org/10.1016/S0166-4115(99)80040-5

Degen, J., & Tanenhaus, M. K. (2019). Constraint-based pragmatic processing. In C. Cummins & N. Katsos (Eds.), *The oxford handbook of experimental semantics and pragmatics* (pp. 21–38). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198791768.013.8

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of the international conference on speech and language processing, Philadelphia* (pp. 1393–1396). IEEE. https://doi.org/10.1109/ICSLP.1996.607874

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P., Friesen, W. V., & Elisworth, P. (1982). What are the relative contributions of facial behavior and contextual information to the judgment of emotion? In P. Ekman (Ed.), *Emotion in the human face* (2nd ed., pp. 111–127). Cambridge University Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *The facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press.

Enos, F., & Hirschberg, J. (2006). A framework for eliciting emotional speech: Capitalizing on the actor's process. In *Proceedings of LREC workshop on emotional speech*. Columbia University Libraries. https://doi.org/10.7916/D88S4ZCG. https://academiccommons.columbia.edu/doi/10.7916/D88S4ZCG

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley Series in Probability and Statistics.

Flexas, A., Rosselló, J., Christensen, J. F., Nadal, M., Olivera La Rosa, A., & Munar, E. (2013). Affective priming using facial expressions modulates liking for abstract art. *PLoS One*, 8(11), e80154. https://doi.org/10.1371/journal.pone.0080154

Garrido-Vásquez, P., Pell, M. D., Paulmann, S., & Kotz, S. A. (2018). Dynamic facial expressions prime the processing of emotional prosody. *Frontiers in Human Neuroscience*, 12, 244. https://doi.org/10.3389/fnhum.2018.00244

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40, 189–212. https://doi.org/10.1016/S0167-6393(02)00082-1

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The karolinska directed emotional faces: A validation study. *Cognition and Emotion*, 22(6), 1094–1118. https://doi.org/10.1080/02699930701626582

Hagoort P. (2019). The neurobiology of language beyond single-word processing. *Science*, 4(6461), 55–58. https://doi.org/10.1126/science.aax0289

Hamann, S., & Canli, T. (2004). Individual differences in emotion processing. *Current Opinion in Neurobiology*, 14(2), 233–238. https://doi.org/10.1016/j.conb.2004.03.010

House, D., Beskow, J., & Granstrom, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. *Proceedings of Eurospeech*, 2001, 387–390. https://doi.org/10.21437/Eurospeech.2001-61

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, 31(1), 19–31. https://doi.org/10.1080/23273798.2015.1072223

Jessen, S., & Kotz, S. A. (2013). On the role of crossmodal prediction in audiovisual emotion perception. *Frontiers in Human Neuroscience*, 7, 369. https://doi.org/10.3389/fnhum.2013.00369

Jessen, S., & Kotz, S. A. (2015). Affect differentially modulates brain activation in uni-and multisensory body-voice perception. *Neuropsychologia*, 66, 134–143. https://doi.org/10.1016/j.neuropsychologia.2014.10.038

Jun, S. A., & Bishop, J. (2015). Priming implicit prosody: Prosodic boundaries and individual. *Language and speech*, 58(4), 459–473. https://doi.org/10.1177/0023830914563368

Jun, S. A., & Fougeron, C. (2000). A phonological model of French intonation. In A. Botinis (Ed.), *Intonation. Text, speech and language technology* (Vol. 15). Springer. https://doi.org/10.1007/978-94-011-4317-2_10

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5), 770. https://doi.org/10.1037/0033-2909.129.5.770

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6), 349–353. https://doi.org/10.1250/ast.27.349

Kim, K. I., Jung, W. H., Woo, C. W., & Kim, H. (2022). Neural signatures of individual variability in context-dependent perception of ambiguous facial expression. *NeuroImage*, 258, 119355.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska directed emotional faces—KDEF*. Karolinska Institute, Department of Clinical Neuroscience, Psychology Section. https://doi.org/10.1037/t27732-000

Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge: MIT Press.

Massaro, D.W. & Cohen, M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, 122(1), 115–124.

Massaro, D. W., & Beskow, J. (2002). Multimodal speech perception: A paradigm for speech science. In B. Granstrom, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp. 45–71). Kluwer Academic Publishers.

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review*, 3(2), 215–221. https://doi.org/10.3758/BF03212421

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748. https://doi.org/10.1038/264746a0

Mobbs, D., Weiskopf, N., Lau, H. C., Featherstone, E., Dolan, R. J., & Frith, C. D. (2006). The Kuleshov Effect: The influence of contextual framing on emotional attributions. *Social Cognitive and Affective Neuroscience*, 1(2), 95–106. https://doi.org/10.1093/scan/nsl014

Mullennix, J., Barber, J., & Cory, T. (2019). An examination of the Kuleshov effect using still photographs. *PLoS One*, 14(10), e0224623. https://doi.org/10.1371/journal.pone.0224623

Paulmann, S., Jessen, S., & Kotz, S. A. (2009). Investigating the multimodal nature of human communication: Insights from ERPs. *Journal of Psychophysiology*, 23, 63–76. https://doi.org/10.1027/0269-8803.23.2.63

Paulmann, S., & Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: How much is enough? *Cognitive, Affective, and Behavioral Neuroscience*, 10(2), 230–242. https://doi.org/10.3758/CABN.10.2.230

Paulmann, S., Titone, D., & Pell, M. D. (2012). How emotional prosody guides your way: evidence from eye movements. *Speech Communication*, 54, 92–107. https://doi.org/10.1016/j.specom.2011.07.004

Pell, M. D. (2005). Nonverbal emotion priming: Evidence from the 'facial affect decision task'. *Journal of Nonverbal Behavior*, 29, 45–73. https://doi.org/10.1007/s10919-004-0889-8

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002

Pourtois, G., Debatisse, D., Despland, P. A., & de Gelder, B. (2002). Facial expressions modulate the time course of long latency auditory brain potentials. *Cognitive Brain Research*, 14, 99–105. https://doi.org/10.1016/S0926-6410(02)00064-2

Prieto, P., Pugliesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, 49, 41–54. https://doi.org/10.1016/j.wocn.2014.10.005

Prince, S., & Hensley, W. E. (1992). The Kuleshov effect: Recreating the classic experiment. *Cinema Journal*, 31(2), 59–75. https://doi.org/10.2307/1225144

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, 105(1), 512–521. https://doi.org/10.1121/1.424522

Rivière, E., Klein, M., & Champagne-Lavau, M. (2018). Using context and prosody in understanding irony: Variability amongst individuals. *Journal of Pragmatics*, 138, 165–172. https://doi.org/10.1016/j.pragma.2018.10.006

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Snow, J., & Mann, M. (2013). *Qualtrics survey software: Handbook for research professionals*. Qualtrics Labs.

Sonntag, G. P., & Portele, T. (1998). PURR—A method for prosody evaluation and investigation. *Computer Speech & Language*, 12(4), 437–451.

Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech*, 46(1), 1–22. https://doi.org/10.1177/00238309030460010201

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Terasawa, H., Berger, J., & Makino, S. (2012). In search of a perceptual metric for timbre: Dissimilarity judgments among synthetic sounds with MFCC-derived spectral envelopes. *Journal of the Audio Engineering Society*, 60(9), 674–685.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society B*, 63, 411–423. https://doi.org/10.1111/1467-9868.00293

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracin, D. (2016). On priming action: Conclusions from a meta-analysis of the behavioral effects of incidentally-presented words. *Current Opinion in Psychology*, 12, 53–57. https://doi.org/10.1016/j.copsyc.2016.04.015

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://doi.org/10.1111/j.1541-0420.2011.01616.x

Zhang, H., Chen, X., Chen, S., Li, Y., Chen, C., Long, Q., & Yuan, J. (2018). Facial expression enhances emotion perception compared to vocal prosody: Behavioral and fMRI studies. *Neuroscience Bulletin*, 34 (5), 801–815. https://doi.org/10.1007/s12264-018-0231-9

Cite this article: Petrone, C., Carbone, F., Audibert, N., & Champagne-Lavau, M. (2024). Facial cues to anger affect meaning interpretation of subsequent spoken prosody, *Language and Cognition*, 1–24. https://doi.org/10.1017/langcog.2024.3