Spatiotemporal structure and substates in facial expressions and speech movements

Hélio Clemente José Cuve[1*], Sophie Sowden-Carvalho[2], Jennifer Cook[2]

[1]School of Psychological Science, University of Bristol, Bristol, United Kingdom

[2]School of Psychological Science, University of Birmingham, Birmingham, United Kingdom

[*] Corresponding author

e-mail: helio.cuve@bristol.ac.uk

School of Psychological Science

University of Bristol

12a, Priory Road

Bristol BS8 1TU

United Kingdom

Abstract

Ranging from overt emotion displays to the understated eyebrow raise accompanying speech, facial expressions are indispensable cues in social interaction. Yet, how these inherently dynamic cues contribute to emotion signalling remains only partially understood. One overlooked factor is how facial actions combine and recombine over time to produce nuanced expressions. Moreover, although most face-to-face interactions naturally involve speech, its contribution to facial emotion signalling has been largely neglected. Building on motor control principles and employing a novel data-driven pipeline - combining spatiotemporal dimensionality reduction, dynamic feature extraction, classification and clustering - we investigated the structure and function of facial dynamics in both *Expression only* and *Emotive speech* signals. We show that a few fundamental spatiotemporal patterns reliably differentiated between specific emotion dynamics such as happy, sad and angry as well as between *Expression only* and *Emotive speech* signals. Furthermore, these spatiotemporal components exhibit meaningful substates, which reflect transient dynamics that are diagnostic of emotion content. This underlying spatiotemporal structure represents an efficient encoding strategy, optimising the transmission and perception of emotion and non-emotion facial signals. These findings have implications for understanding normative and "atypical" face-to-face signalling and can inform the design of expressive emotive capabilities in artificial social agents. Moreover, our data-driven pipeline can be extended to study broader aspects of multimodal social signalling.

Keywords: Facial expressions, movement, spatiotemporal, dimensionality reduction, emotion, signalling.

Significance Statement

In face-to-face interactions, facial movements convey both emotional and speech-related information, yet their complex dynamic signalling remains only partly understood. This study explores how facial movements evolve over time, integrate emotional and speech cues. Using advanced techniques to measure and analyse moment-to-moment changes in dozens of facial muscles we identified fundamental movement patterns that differentiate emotional signals. These dynamic patterns enable both isolated emotional signalling and simultaneous emotion cues during speech. These findings can inform understanding of typical and atypical face-to-face interactions abilities and enhance the design of human-like facial expressions in social robots and virtual agents. We also contribute a novel data-driven pipeline for analysing facial dynamics, applicable to broader facets of multimodal social communication.

Introduction

The human face is an incredibly versatile signalling tool for social interaction (1) capable of producing a rich array of non-verbal signals through the coordination of basic motor facial action units (AUs) (2). For example, frowning (AU4) combined with pursed lips (AU24) can convey a prototypical facial expression of anger; while raising the cheeks (AU6) and pulling back the lip corners (AU12) can signal a smile. In addition to ostensive non-verbal emotional cues, in real-life social interaction, facial expressions simultaneously modulate verbal communication. A slight tightening of the eyelids (AU7) for example, can make a “neutral” verbal utterance appear slightly angrier. The mechanisms that afford such flexible expressive capabilities to facial expressions remain only partly understood.

One overlooked aspect in previous research is the dynamic nature of facial expression signalling, with a large focus instead on the static aspects of facial behaviour or approaches ignore temporal dynamics (3). Moreover, although most face-to-face interactions naturally involve speech, the role of speech-related facial dynamics in emotion signalling has been largely neglected. Yet these dynamic cues in conversational or speech-related facial expressions can enhance the simultaneously transmission of visual signals for emotion and verbal communication (3–6). Nonetheless, a few key studies have provided important insights on the role of facial expression dynamics to emotion signalling. Psychophysics inspired approaches have combined generative synthesis of facial movements (7) and reverse correlation to parametrically map the spatiotemporal features of AUs to perceptual representations of emotion categories (8), valence-arousal dimensions (9) and even social traits (10). Similarly, a recent production and perception study recorded participants' facial

movements during signalling of anger, happiness, and sadness, as well as during spoken sentences conveying the same emotions (11). Kinematic analysis revealed faster movements for high-arousal emotions (happiness, anger) compared to slower movements for low-arousal emotions (sadness). Thus, suggesting that facial dynamics carry a diagnostic signal for differentiating emotional expressions at both the production and perception stages.

Nevertheless, a detailed account of the spatiotemporal structure of facial signals—how AUs dynamically combine and unfold to signal emotions and simultaneously convey speech-related information—remains lacking. One significant challenge to studying the role of spatiotemporal structure in facial expression signalling is the lack of an integrated theoretical and methodological framework. Such a framework should ideally integrate the dynamic nature of facial expression signals with their socio-communicative functions and provide an appropriate modelling strategy. In this work, we build on the observation that facial expressions, like other facial and bodily movements, are controlled by an underlying motor system (12, 13). We explore the application of theoretical and analytical principles drawn from the motor control literature to understand the spatiotemporal structure and function of dynamics in facial expressions.

A motor control perspective provides two useful insights that are applicable to studying the spatiotemporal structure and function of facial expression dynamics. First, a longstanding motor control hypothesis suggests that body movements are generated by a low-dimensional mechanism that underpins their effective neural control and physical execution (14–16). Such low-dimensionality of motor control and execution

might reflect the brain's solution to the problem of controlling numerous muscles and components of the motor system whilst minimising computational demands - often referred to as the "degrees of freedom" problem (17–19). Crucially, such low dimensionality has been shown to describe motion patterns during walking (20), whole body pointing (21), eye movements (22) and articulatory speech movements (23). These movements also convey emotion-related information (24–26) suggesting the potential for shared control mechanisms for social and non-social aspects of movement. Similarly, dynamics underlying facial expressions, and their ability to multiplex information (e.g., visually signal emotion and speech), may be structured around flexible low-dimensional spatiotemporal patterns. Such low dimensional patterns could, for example, reflect how AUs combine and evolve over time during facial expression production and simultaneously encode emotion in speech-related facial movements. This perspective also suggests a methodological framework to model spatiotemporal nature of movement, which is rare in the study of facial expression production. For instance, this can be explored using data-driven models that learn a reduced dimensionality from high-dimensional AU movement data (27).

A second insight from the motor-control literature is the observation that although outwardly fluid, body movements can often be segmented into distinct 'substates' often seen in differential modulation of kinematics (e.g., speed, displacement of motor actions). For instance, walking movements can be segmented into 'swing' or 'stance' phases (20) or eye movements into 'fixations' and 'saccades' (28), and even more fine-grained states (sub movements) (29). Facial expressions are likely to have substates related to the biomechanics of how they are produced, encompassing 'transitions'— where AUs are initiated or inhibited, - and 'stability' – where AUs are sustained (30,

31). However, to date facial expressions substates of facial expressions and speech movements have not been formally characterised, and whether and how they relate to emotional signalling is unknown. Similar to spatiotemporal components, substates can be explored in a data-driven fashion, using methods such as spatiotemporal clustering (27, 32, 33) without *apriori* defined kinematic profiles.

Some prior work has indeed proposed and employed similar motor-centric theoretical and methodological approaches to account for and leverage the rich dynamics of facial signals in perception and production of emotion cues. Arguably, the majority and significant contributions in this area have focused on perception e.g. - (8, 34) rather than production which is the focus of the current study. Notwithstanding, a few recent studies have employed motor-informed spatiotemporal dimensionality reduction approaches to identify dynamic motifs in facial expression production data (e.g. (18, 35, 36). Nonetheless, the reliance on only single to a handful signallers or actors arguably limits the dynamic range and variability the facial signals analysed. Hence, it's unclear whether the observed low dimensionality in these studies, genuinely reflect the underlying mechanisms for facial expression signalling or simply results from restricted variability in signal dynamics. It is also unclear how well these findings generalise to other emotional expressions and conditions (e.g., emotive speech facial dynamics), and signallers. Furthermore, the majority of these studies still relies on averaging across the spatiotemporal dimensions when comparing expressions and conditions, potentially overlooking how inherent dynamic patterns relate to emotion signalling (see Methods). Consequently, the functional role of facial dynamics—

including both low-dimensional spatiotemporal patterns and substates—in emotional signalling remains only partially explored.

In contrast, probing the underlying spatiotemporal dimensions and substates in facial expressions enables a richer characterisation of facial dynamics that is both holistic, i.e., captures overall spatiotemporal patterns, but also focal as it can capture micro spatiotemporal patterns in movement data. Crucially, we hypothesise that these underlying patterns are not merely incidental; rather, they are diagnostic of emotional content, reflecting how facial movements are optimised for social signalling.

The current study therefore aims to (1) describe the spatiotemporal structure underlying facial expression production; (2) explore and characterise potential facial expression substates, and 3) evaluate the diagnostic value of spatiotemporal structure and substates for emotion signalling. To capture wider facial dynamics relevant for social interaction, we recorded participants facial movements while they signalled different emotions in a) *Expression only* and b) *Emotive speech* conditions across two separate sessions. We used automated Facial Action Coding System (FACS) to describe the spatiotemporal patterns of the facial AUs (e.g., mouth widening, eyebrow raise, etc). We then employed a novel data-driven pipeline combining dimensionality reduction, dynamic feature extraction, and spatiotemporal classification and clustering; to identify diagnostic latent spatiotemporal patterns (groups of co-activated AUs and salient temporal patterns) underlying optimal emotion signalling in both Expression only and Emotive speech facial dynamics.

To precis our results, we find that *Expression only* and *Emotive speech* facial dynamics can be distilled into a of small set of fundamental spatiotemporal

185 components that form the basis for flexible signalling. Specifically, the three emotions
186 (happy, angry, sad) and two conditions (*Expression only* and *Emotive speech*) studied
187 are differentiated by synergistic mixing of components and subtle local substate
188 differences related to expression transitions. These results suggests that the complex
189 and flexible signalling of emotional through facial movements, is achieved through
190 variations in a small number of spatiotemporal components and substates.
191

192 ## Results

193 **Spatiotemporal components for facial expression production.**

194 We report results for *Expression only* and *Emotive speech* expressions separately for
195 clarity, however we note that these patterns were consistent when analysed together.
196

197 Spatiotemporal patterns during facial *Expression only* condition

198 We first looked at the spatiotemporal components identified via the Non-Negative
199 Matrix Factorization (NMF), which represent the underlying (reduced) spatiotemporal
200 patterns (e.g., groups of AUs with common activation timecourses and salient
201 temporal patterns). The optimal number of spatiotemporal components (*k*) optimising
202 for various fit indices (see Methods) – was consistently three (see Fig 1 – top panel)
203 (see also *Expression only* condition - NMF validation analysis in Supplementary
204 Information – Fig S1-3).

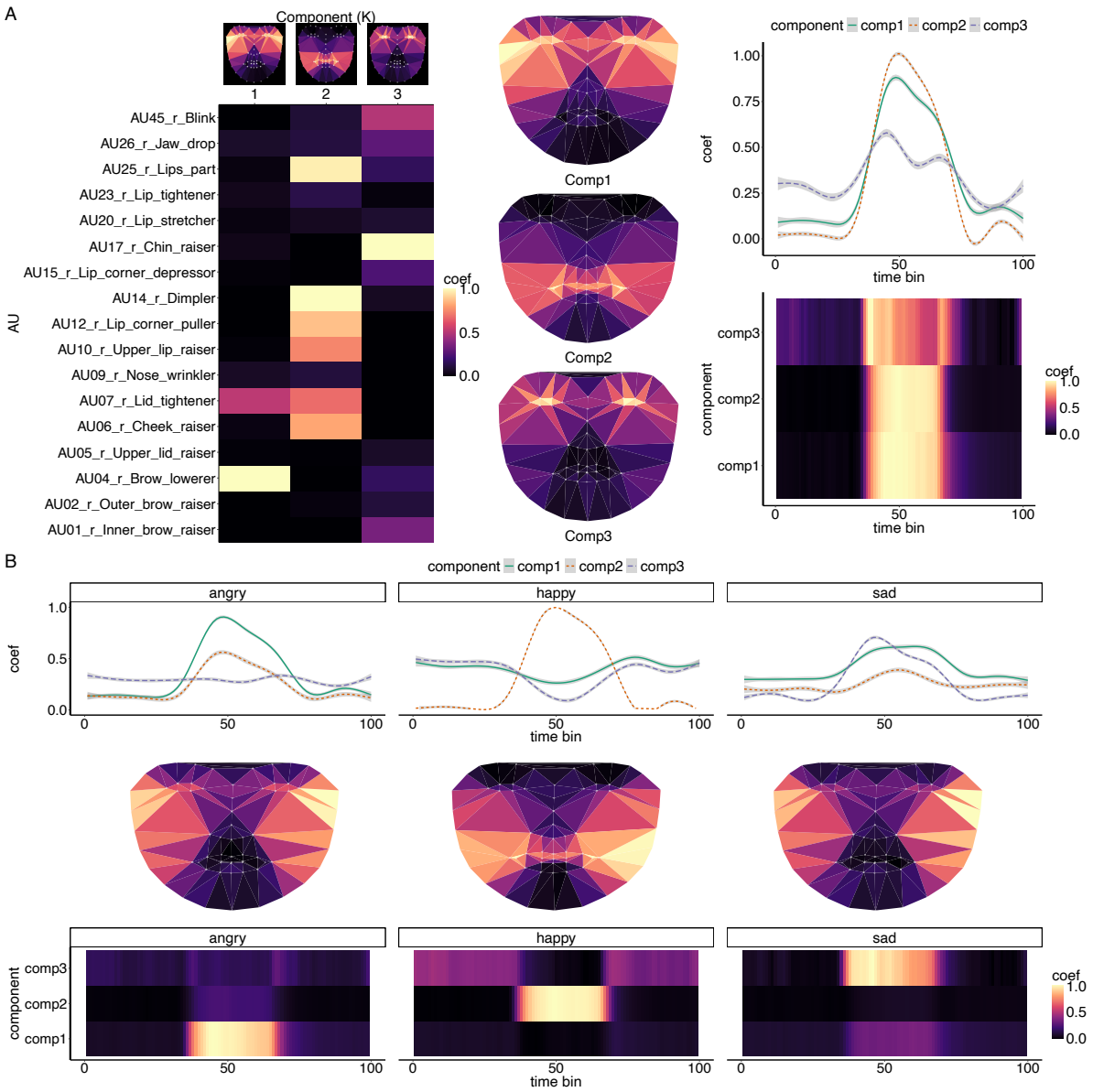

Figure 1. Spatiotemporal components for facial expressions only production. The top panel (A) shows that expression production can be summarised by three components, capturing spatiotemporal patterns for upper face AUs (component 1), lower face AUs (component 2) and a combination of both lower and upper face AUS (component 3). Top - middle and right columns show approximate spatial distribution of components in an average face model and how it changes over time, respectively. The bottom panel (B) highlight spatiotemporal profiles for angry, happy and sad expressions. Coef = coefficient from the NMF, brighter colours indicate stronger activation. Note while heatmaps are drawn to scale, face maps are drawn for visualisation purposes only.

Evaluating the spatiotemporal patterns revealed that component 1 (C1) had stronger activation of the upper face-related AUs (e.g., AU4 brow lowerer, AU7 lid tightener)

compared to other AUs. Component 2 (C2) had strong activation of both lower and upper face-related AUs (e.g., AU25 lips part, AU14 dimpler). Component 3 (C3) was mostly dominated by lower face actions (e.g., AU17 chin raiser) with some small activation of upper face AUs (e.g. AU45 blink). We name these components, upper face AUs (C1), lower face AUs (C2) and lower-to-upper face AUs (C3). The temporal profile of these components is shown in Fig 1 B. C1 (upper face AUs) and C2 (lower face AUS) had a common inverted-U shape with a clear singular peak, whereas C3 (lower to upper face AUs) had a less pronounced but more dynamic pattern (e.g. bimodal peaks).

Next, we assessed the diagnostic value of the spatiotemporal components, by analysing how their dynamic features differed between expressions and validated it in a held-out test set (see Methods - *Multivariate Timeseries Classification*). When assessing spatiotemporal components' ability to classify expressions we achieved significantly above chance accuracy of over 90% (ACC (accuracy)= 0.91 [BalACC (balanced accuracy) = .93], Kappa = 0.87; AUC(Area Under the Curve) mean = 0.98, $p < .001$ see Fig 2). These results indicate strong agreement between prediction and ground-truth – see also SI1 – Table S1-2

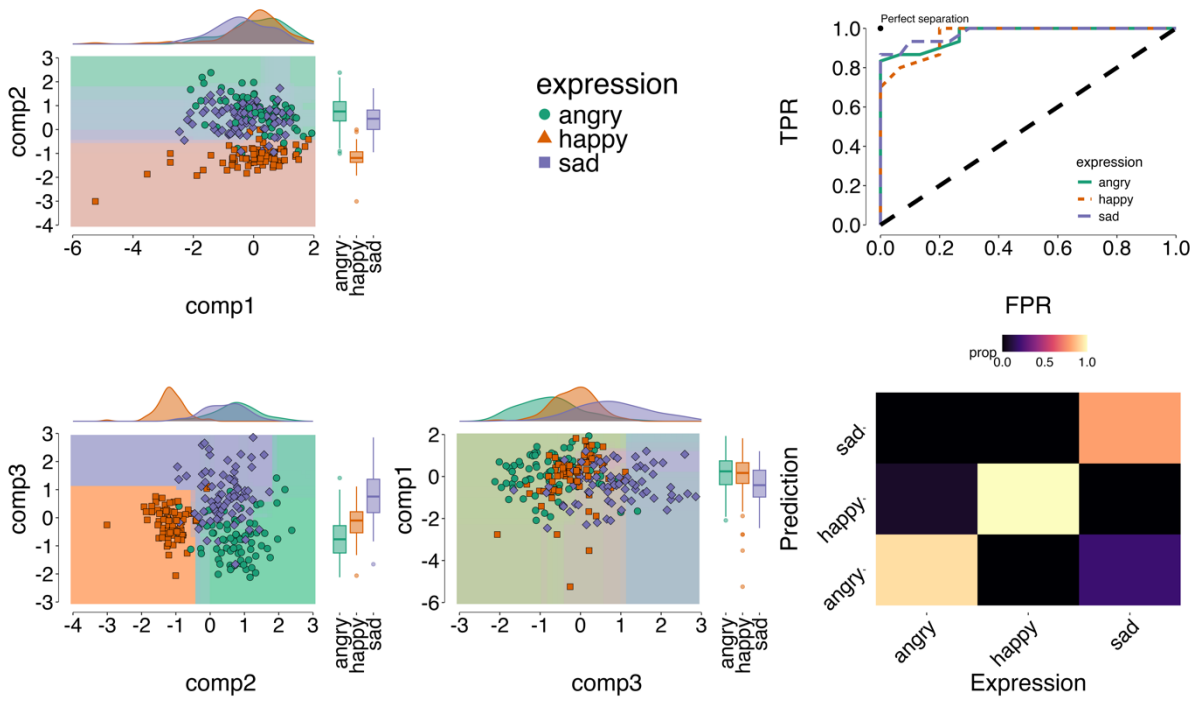

Figure 2. Spatiotemporal classification of facial expressions only production. *Notes:* The left panels show decision boundaries (filled backgrounds) across spatiotemporal components (Comp1: upper AUs, comp2: lower face AUs; comp3: lower to upper face AUs) estimated on a training set to predict emotion classes (angry, happy, sad). Coloured dots represent individual ground-truth expressions from both training and held-out test data. Dots within a matching background indicate correct classifications, while mismatches indicate misclassifications. The results demonstrate robust classification, with at least one component combination separating an expression from the rest. e.g. comp1 vs 2 and comp-2 vs 3 robustly distinguishes happy expressions. The top right panel presents Receiver Observer Characteristics (ROC) curves, showing good model performance (higher True Positive Rate [TPR] and lower False Positive Rate [FPR]) with the diagonal dashed line representing chance). The bottom right panel displays a confusion matrix. Brighter colours indicate higher accuracy along the diagonal elements, with off-diagonal showing little to no confusions (darker colours).

We also found distinct temporal profiles for each emotion. Angry expressions (BalACC = .91, AUC=.97, p < .001) were characterised by strong temporal coactivation of C1 (upper face AUS) and C2 (lower face AUS). Happy expressions (BalACC = .98, AUC = .99, p < .001) were dominated by C2 (lower face AUs); and sad (BalACC = .98, AUC = .90, p<.001) characterised by moderate changes in all three components. To further validate our model, we conducted a Fisher's exact test on the confusion matrix (see

Methods), which confirmed significant pairwise differentiation of all expressions on the basis of their spatiotemporal patterns (angry vs happy; angry vs sad; and happy vs sad: p.corrected < .001). These results reinforce the robustness of our classification model, demonstrating that spatiotemporal patterns differ between emotional expressions.

Spatiotemporal patterns for *Emotive speech* facial expressions

Similar to the *Expression only* condition, we found that three components effectively represented the spatiotemporal patterns of facial AUs in *Emotive speech* expressions (see SI1 – Fig S4-6). However, there were some notable differences in the spatial and temporal patterns, which likely reflect the fact that, in contrast with the *Expression only* videos, the *Emotive speech* videos comprise a combination of facial expression and speech AUs. Specifically, the *Emotive speech* components comprised a mix of both lower and upper face AUs, in comparison to mostly lower vs upper face AUs for *Expression only* conditions (see Fig 3).

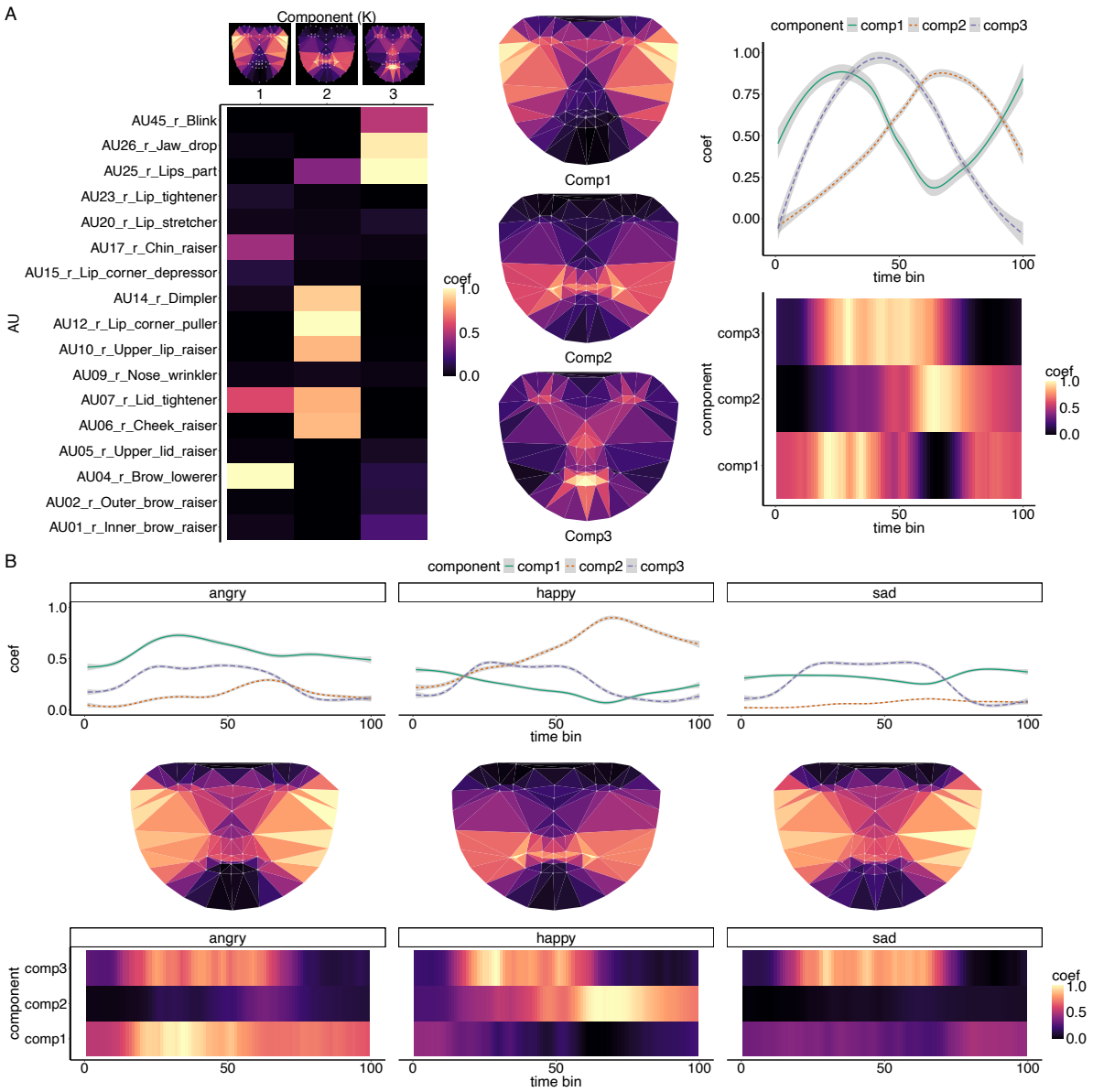

Figure 3. Spatiotemporal components for *Emotive speech* facial expression production. The top panel (A) shows that *Emotive speech* facial expression production can be summarised by three underlying components, capturing synergies between upper to lower face AUs (Component 1), lower-mid-upper AUs (component 2) and lower to upper face AUs (component 3). Top - middle and right columns show approximate spatial distribution of components in an average face model and how it changes over time, respectively. The bottom panels (B) show how components change over time, highlighting different profiles for angry, happy and sad expressions. Coef = coefficient from the NMF, brighter colours indicate stronger activation. Note while heatmaps are drawn to scale, face maps are not a one-to-one map and are drawn for visualisation purposes only.

Component 1 included strong activation of upper and lower face Aus (e.g., AU4 - brow lowerer; AU17 - chin raiser) peaking at both early and late time points. Component 2

(lower-mid-upper face AUs) combined strong activation of AUs in the lower (e.g., AU25 - lips part), middle (e.g., AU14 - dimpler) and upper face areas (e.g., AU7 - lid tightener). Component 3 primarily featured activation of lower face AUs, such as AU25 (lips part) and AU26 (jaw drop), plus additional moderate activation for AU 45 (blink). Henceforth component 3 is referred to as lower to upper face AUs (see Fig 3). C2 (lower-mid-upper face AUS) and C3 (lower to upper face AUs) had an inverted-U shape with C3 peaking earlier than C2 (Fig 3C).

As with facial *Expression only* production, *Emotive speech* expressions were also characterised by distinct spatiotemporal patterns. That is, a classifier trained on timeseries-based features of the spatiotemporal components (e.g., curvature, entropy) achieved significant accuracy in classifying expressions on a held-out test set, with performance at 70% (ACC = 0.69, 95% CI: 0.53, 0.81; BalAcc = .76 Kappa = 0.53, $p < 0.001$). These results indicate moderate to strong agreement between prediction and ground-truth emotional expression labels.

Class-wise, the spatiotemporal components significantly distinguished all expressions. Happy had the highest accuracy (BalACC = 0.91, AUC = 0.97, $p < .001$), suggesting a clear and distinct spatiotemporal profile, notable in the increased activation in C2 (lower-mid-upper face AUs) over time (see Fig 4).

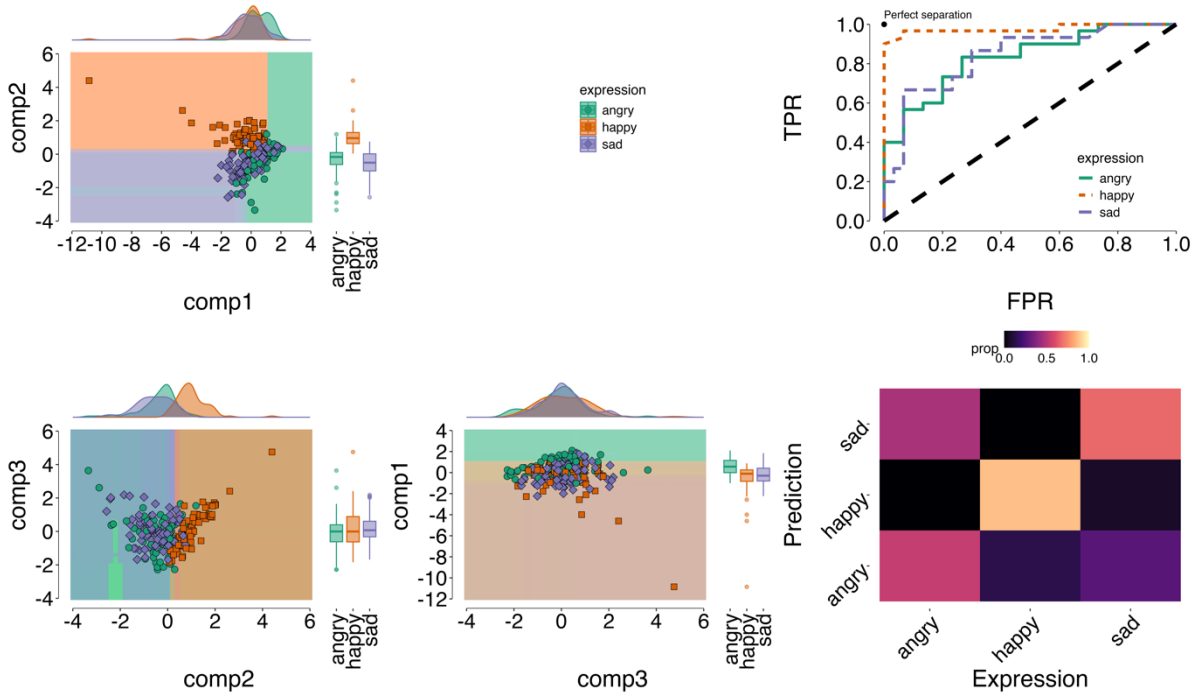

Figure 4. Spatiotemporal classification of *Emotive speech* facial expressions. The left panels show decision boundaries (filled backgrounds) across spatiotemporal components (comp1: upper to lower face AUs, comp2: lower to upper face AUs; comp3: upper-to-lower face AUs) estimated on a training set to predict emotion classes (angry, happy, sad). Coloured dots represent individual ground-truth expressions from both training and unseen test data. Dots within a matching background indicate correct classifications, while mismatches indicate misclassifications. The results demonstrate robust classification, with component combinations effectively separating the expressions in at least one combination. The top right panel presents ROC curves, showing good model performance (higher True Positive Rate [TPR] and lower False Positive Rate [FPR]), with the diagonal dashed line representing chance. The bottom right panel displays a confusion matrix. Brighter colours indicate higher accuracy along the diagonal elements, and confusion off-diagonal (e.g., happy shows high accuracy and low confusion, whereas angry shows higher confusion than all expressions).

Angry had the lowest classification accuracy compared to all other expressions, yet still significantly above chance (BalACC = 0.67, AUC = 0.83, p = .01), reflecting a more complex and less distinct pattern across all components. Similar to happy, sad expressions were classified reliably above chance (BalACC = 0.72, AUC = 0.84, p = 0.001). Fisher's exact test on the confusion matrix confirmed significant pairwise differentiation between happy and sad expressions (p < .001) and between angry and

happy ($p < .001$). However, it suggested confusions between and angry and sad ($p = 0.12$) (see Fig 4; SI1 - Table S3-4).

In summary, we found three latent spatiotemporal components that reliably described the dynamics of facial expressions during *Expression only* and *Emotive speech* conditions. Robust train-test classification on unseen held-out data, indicates that spatiotemporal components carry information that is diagnostic of emotional content.

Substates reveal emotion diagnostic transitory patterns and sequences in facial signals

The second objective of this study was to identify and characterise facial expression substates, such as transition and sustain periods, and their diagnostic value for emotion signalling (i.e., whether they differ between emotions: happy, angry and sad expressions). We identified three primary substates through spatiotemporal clustering that segmented periods with robust local dynamics in the spatiotemporal components (see Methods: *Spatiotemporal clustering for substates analyses*). Visual representation of the kinematic profiles of each substate (Fig 5A-B) suggest the substates correspond to relaxed, transition, and sustain periods for spatiotemporal components. The overall pattern of substates is also consistent across participants (see Fig 5 C-D).

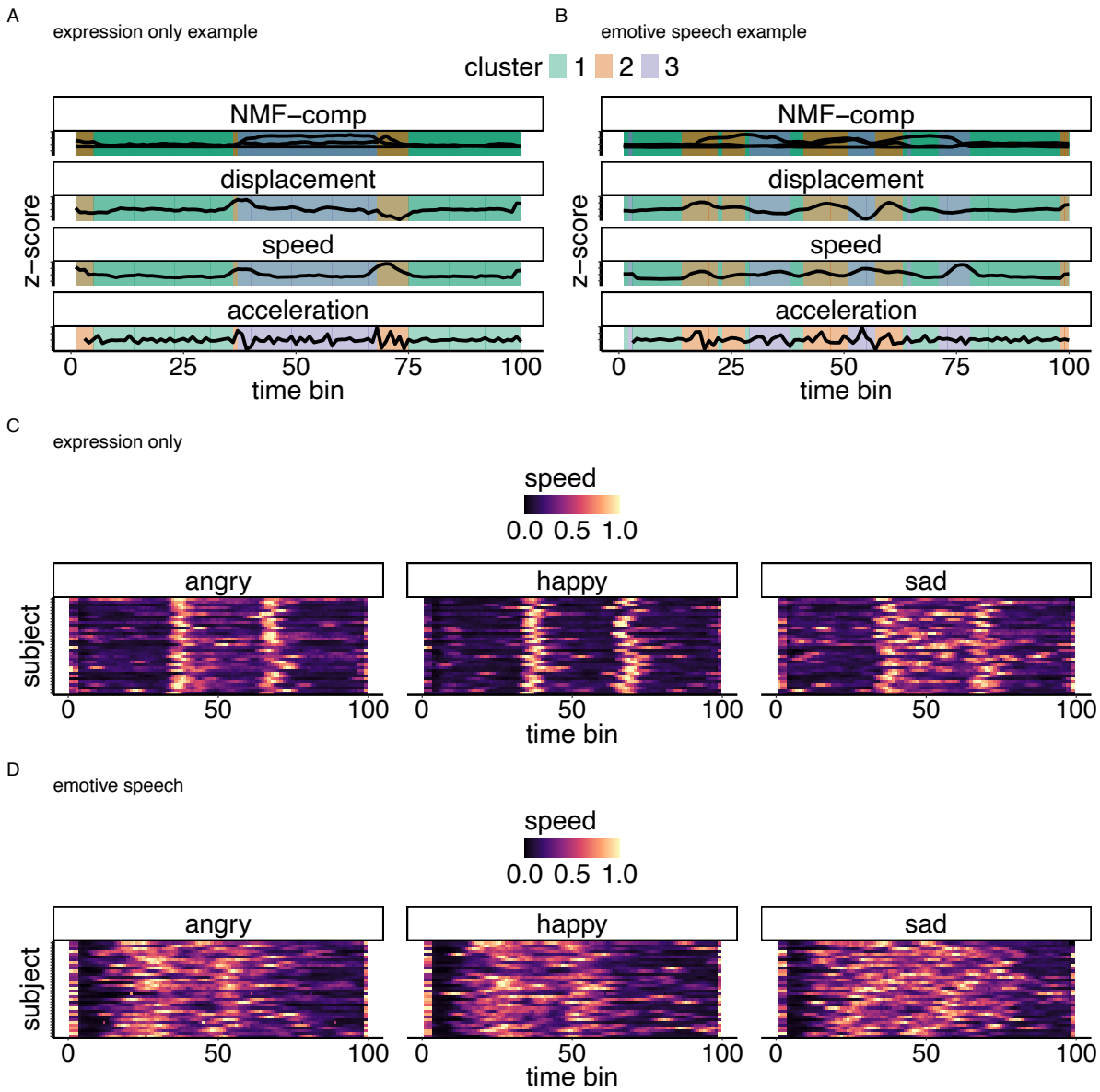

Figure 5. Illustration of facial expression substate profiles.
A-B. Example of substate clusters (1 (green) = relaxed, 2 (peach) = transition; 3 (purple) = sustain). C-D show overall the speed distribution for *Expression only* and *Emotive speech* expressions for all subjects.

To illustrate the pattern of facial expression substates and how it differs between emotions (anger, sadness and happy) and conditions (*Expression only vs Emotive speech*), we computed two metrics: (1) speed within the substate hence providing a kinematic characterisation of each substate, and (2) entropy, which provides a view of overall dynamics of substates (e.g., how they change over time) (Fig 5) – see also

Methods. Similar measures have been shown to capture diagnostic information about facial expression movement patterns (8, 11).

Transition substate speeds differentiate emotional expressions

A mixed-effects model (see Methods) predicting speed from emotion (happy, angry, sad), condition (*Expression only, Emotive speech*), and substate (relaxed, transition, sustain) showed strong total explanatory power (conditional R2 = 0.6 (66%)) with the fixed effects alone (marginal R2) accounting for 0.63 (63%) of the variance. All main effects were significant. First, there was a small significant effect of emotion ($F(2,1990) = 12.89$, $p < .001$; $\eta_p^2 = 0.02$, 95% CI [0.009, 0.10]). Pairwise tests showed that angry expressions were significantly slower than happy expressions (angry < happy; beta = -0.028, SE = 0.008, $t(1207) = -3.398$, $p = 0.002$). Happy expressions were faster than sad expressions (happy > sad; beta = 0.041, SE = 0.008, $t(1207) = 4.92$, $p < 0.001$). There were no significant differences in speed between angry and sad expressions (beta = 0.012, SE = 0.008, $t(1207) = 1.53$, p.corrected = 0.38). There was a significant main effect of condition ($F(1, 1190) = 369.41$, $p < .001$; $\eta_p^2 = 0.24$, 95% CI [0.20, 1.00]) with *Expression only* condition moving faster than spoken (beta = 0.13, SE = 0.007, $t(1207) = 19.08$, $p < .001$). Finally, there was a main effect of substate ($F(2, 1190) = 687.90$, $p < .001$; $\eta_p^2 = 0.54$, 95% CI [0.51, 1.00]) where sustain had higher speed than relaxed (beta = -0.16, SE = 0.008, $t(1207) = -19.52$, p.corrected < .001); transition had higher speed than relaxed (beta = -0.307, SE = 0.008, $t(1207) = -36.80$, p.corrected < .001) and transition had higher speed than sustain (beta = -0.144, SE = 0.008, $t(1207) = -17.28$, p.corrected < .001).

All three two-way interactions were significant. First, there was a significant interaction between emotion and substate ($F(4, 1190) = 15.24$, $p < .001$; $\eta_p^2 = 0.05$, 95% CI [0.03, 1.00]). Pairwise tests showed that differences in speed between happy, angry and sad emotional expressions were statistically significant for the transition substate only (happy>anger, beta = -0.077 SE = 0.014, t(1207) = -5.36, p.corrected < .001; sad< angry, beta = 0.053, SE = 0.014, t(1207) = 3.64, p.corrected < 0.001; sad< happy, beta = 0.130, SE = 0.014, t(1207) = 9.005, p.corrected < .001). There were no speed differences between emotions for the relaxed substate (angry vs. happy, beta = -0.005, SE = 0.014, t(1207) = -0.348, p.corrected = 1.0; angry vs. sad, beta = 0.007, SE = 0.014, t(1207) = 0.498, p.corrected = 1.0; happy vs. sad, beta = 0.012, SE = 0.014, t(1207) = 0.846, p.corrected = 1.0) nor the sustain substate (angry vs. happy, beta = -0.003, SE = 0.014, t(1207) = -0.177, p.corrected = 1.0; angry vs. sad, beta = -0.022, SE = 0.014, t(1207) = -1.501, p = 0.40, p.corrected; happy vs. sad, beta = -0.01913, SE = 0.0144, t(1207) = -1.325, p.corrected = 0.5567). These results suggest that transition substate kinematics (but not relaxation or sustainment kinematics) carry diagnostic information about the emotional content of facial expressions (Fig 6).

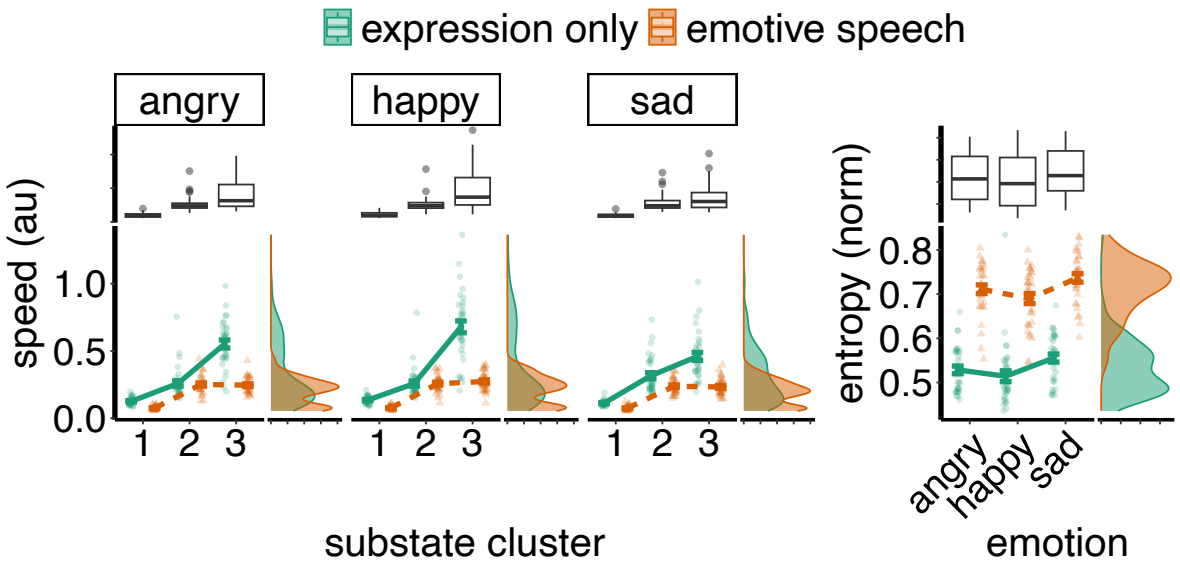

Figure 6: Summary of speed and entropy of substate patterns.

Substate cluster: 1-relaxed, 2-sustained, 3-transition. *Expression only* condition has faster than *Emotive speech*, with transition substates being faster; Entropy mainly differentiates *Expression only* and vs. *Emotive speech* condition, with the latter exhibiting more complex patterns. Entropy also differed between happy and all other expressions, with substate patterns being more predictable for former.

In addition, there was a significant two-way interaction between emotion and condition (F(2) = 3.54, p = 0.029; η_p^2 = 0.006, 95% CI [3.15e-04, 1.00]). For the *Expression only* condition, there were significant speed differences between two out of the three emotion expression pairs: specifically, happy > angry (beta = -0.045, SE = 0.012, t(1207) = -3.84, p.corrected < 0.001); happy > sad (beta = 0.06171, SE = 0.0118, t(1207) = 5.232, p.corrected < .001); and no differences between angry vs. sad (beta = 0.016, SE = 0.012, t(1207) = 1.39, p.corrected = 0.492). In contrast, for the *Emotive speech* condition, there were no significant differences in speed between the emotional expressions (angry versus happy, beta = -0.011, SE = 0.012, t(1207) = -0.97, p.corrected = 1.0; angry versus sad, beta = 0.009, SE = 0.012, t(1207) = 0.77, p.corrected = 1.0; happy versus sad, beta = 0.020, SE = 0.012, t(1207) = 1.73, p.corrected = 0.25).

Finally, there was a significant two-way interaction between condition and substate (F(2) = 183.65, p < .001; η_p^2 = 0.24, 95% CI [0.20, 1.00]). For the *Expression only* condition, there were significant differences in speed between all substates such that sustain was faster than relaxed (relaxed < sustain, beta = -0.153, SE = 0.012, t(1207) = -12.96, p.corrected < .001); transition was faster than relaxed (relaxed < transition, beta = -0.439, SE = 0.012, t(1207) = -37.23, p.corrected < .001), and faster than sustain (beta = -0.286, SE = 0.0118, t(1207) = -24.271, p.corrected < .001). These patterns were overall similar for the *Emotive speech* condition where relaxed periods were significantly slower than sustain (beta = -0.173, SE = 0.0118, t(1207) = -14.647,

p.corrected < .001) and transition periods (beta = -0.175, SE = 0.012, t(1207) = -14.82, p < .001). However, there was no statistically significant difference between sustain and transition speeds (beta = -0.002, SE = 0.012, t(1207) = -0.169, p.corrected = 1.00). In sum, speed analysis confirm that subset clustering is associated with differences in kinematics, and these also differs between expressions and between *Expression only* and *Emotive speech* signals.

<u>Substate dynamics differentiate emotion content</u>

Here we used entropy as a metric to statistically describe the dynamics of facial expression substates in terms of how temporally structured/predictable they are ((37). see Methods). High entropy equates with high complexity in substate patterns, and low entropy with low complexity. Fig 5, for example, illustrates that substate transitions are more complex (i.e., high entropy, less predictable) in the *Emotive speech* condition where visual pattern differences between emotions are harder to discern than in the *Expression only* condition (see Methods: *Spatiotemporal clustering for substates analyses*).

The entropy mixed model explained 67% of the variance ($R^2 = 0.67$). The fixed effects alone (emotion, and expression condition) accounted for 57% of the variance (marginal $R^2 = 0.57$). The main effect of emotion was statistically significant ($F(2, 340) = 12.27$, $p < 0.001$; $\eta p^2 = 0.07$, 95% CI [0.03, 1.00]). Pairwise contrasts revealed that sad expressions had a significantly higher entropy (more unpredictable substate sequences) compared to happy expressions (estimate = -0.027, SE = 0.0055, t(344) = -4.897, p < 0.0001). There were no statistically significant differences in entropy between angry and happy expressions (estimate = 0.011, SE = 0.006, t(344) = 1.996,

$p = 0.14$), but there was a significant difference between angry and sad expressions (estimate = -0.016, SE = 0.0055, $t(344) = -2.90$, $p = 0.012$).

The largest effect was the main effect of condition ($F(1, 34) = 429.10$, $p < 0.001$; $\eta p^2 = 0.93$, 95% CI [0.89, 1.00]), with the *Expression only* condition demonstrating significantly lower entropy than *Emotive speech* expressions ($\beta = -0.208$, SE = 0.017, $t(35) = -12.594$, $p < 0.001$). This suggests more complex movement dynamics in the *Emotive speech* compared to the *Expression only* condition (see heatmaps in Fig 5). There was no significant interaction between emotion and condition ($F(2, 340) = 0.03$, $p = 0.970$; $\eta p^2 = 0.0002$, 95% CI [0.00, 1.00]). In summary, *Expression only* substates are more temporally structured than those in *Emotive speech* expressions. Nonetheless, substates patterns differentiate expressions with sad expressions having a more complex pattern of substates than happy and sad expressions.

To summarise, in addition to the low-dimensional spatiotemporal patterns that are diagnostic of emotion and expression condition, we also identified three substates reflecting periods of relaxation, transition, and sustainment of facial actions. These substates are distinguished by their unique kinematic and dynamic (complexity) profiles which differ between emotional expressions and *Expression only* versus *Emotive speech* conditions.

Discussion

The current study explored how the underlying dynamics of facial movements relate to their emotional signalling function. Building on a motor-theoretic inspired approach we employed a novel data-driven pipeline - integrating spatiotemporal dimensionality

reduction, timeseries feature extraction, classification and clustering; to identify diagnostic patterns underlying emotion signalling in both *Expression only* and *Emotive speech*. We consistently identified a small number of components that reliably described the time courses and covariance patterns of over a dozen action units. The identification of these components suggests that emotion signalling through facial expressions and speech-related cues is achieved by just a few fundamental dynamic patterns; combinations of these dynamic patterns differentiate angry, happy and sad expressions. Second, we showed that these spatiotemporal patterns exhibit substate-like structure, which reflect transient periods of relaxation, transition, and sustainment of expressions. The kinematic profiles of these substates differed by emotion category as well as between *Expression only* and *Emotive speech* conditions.

The low dimensionality of the spatiotemporal patterns of facial expressions, that we have uncovered, is intriguing given the various combinations of AUs that can contribute to the same expressions (3) as well as the variability in how people produce expressions (38). A low-dimensional spatiotemporal structure may reflect flexibility for signalling, as a variety of expressions can be rapidly produced by relying on just a few interchangeable dynamic patterns. Such flexibility is likely beneficial in face-to-face interaction, where facial expressions may need to be quickly adapted according to interaction demands (3, 39) or where it is necessary to accentuate emotional information during speech (4, 40). These findings are also consistent with the observation of low-dimensional spatiotemporal patterns facial (18, 36) as well as other body movements (17, 19, 21) which also have the capacity to signal emotion information (24, 41, 42). Thus, it suggests the potential for shared control mechanisms for social and non-social aspects of movement including speech movements.

An important question concerns whether our results are representative of facial expression signalling in general, or simply an outcome of the particular methodology we employed. For example, it could be argued that we identified a small number of spatiotemporal components because we asked participants to demonstrate relatively simplistic expressions which can be signalled by activating just a few action units. We argue that this is not the case because a) the *Emotive speech* condition was also specifically designed to dissuade participants from relying on overly simple and caricatured expressions and to introduce dynamic variability and b) compared to previous work (18, 36) we provide a substantially wider range of signallers and hence capturing wider idiosyncrasies and variability in dynamics. It could also be argued that our results are simply a consequence of dimensionality reduction, as any set of data can in theory be represented by a lower-dimensional structure. Yet, as we showed in a series of validation and sensitivity analyses, the components learned carry diagnostic information about emotions, which is robust to comparisons against appropriate randomised dynamic patterns and generalises to unseen data. Thus, we have not generated a “non-sense” low dimensional structure but rather a meaningful way to represent the data.

An intriguing question concerns why this low-dimensional spatiotemporal structure emerges. One possibility is that receivers only attend to a reduced subset of spatiotemporal patterns and so, correspondingly, signallers tend to only move a reduced subset of features, or to move them in predictable ways. In support of this, studies have found that even when observers are presented with dozens of biologically plausible face movements, their perceptual representations suggest that they only rely

on a reduced subset (34). Since human infants learn to produce and recognise facial expressions throughout a protracted period of development (43), it is possible that humans learn to optimise signals based on how these are optimally perceived and responded to, creating a reciprocal dynamic interaction between these processes. Thus, the low-dimensional organisation of facial expressions may reflect the optimisation of social signals according to perceptual demands and these perceptual demands, themselves, may be influenced by the low-dimensional organisation of facial expressions (44). An alternative possibility is that there are savings (e.g., energy/computational) to be had from moving the face according to a few spatiotemporal patterns.

The identification of substates in emotional facial dynamics also goes beyond previous work on facial expression production. On the one hand substates likely reflect the biomechanics of producing facial expressions, which inevitably result in periods of relaxation, transition and sustainment of facial motor actions (45). Yet, the fact that the patterns of these substates vary with different emotions and expressive conditions, suggests that the local dynamics are sensitive to the communication of different emotions.

In real-life interaction, facial movements subserving verbal and nonverbal expressions are often seamlessly integrated. Our findings in the *Emotive speech* condition raise the possibility that verbal and nonverbal face dynamics when signalling emotion are integrated by modulating different spatiotemporal components. For example, during speech, lip movement might covary with articulation and sound production while simultaneously, the tightening of the lips and lowering the eyebrows can accentuate

556 the emotion signal. Indeed, the observation of mostly part-based profiles (e.g., lower
557 vs upper face AUs) in *Expression only* conditions, contrasting the involvement of both
558 mouth and non-mouth-related components in *Emotive speech* expressions supports
559 this assertion. Interestingly, the higher entropy observed in the substate patterns of
560 *Emotive speech* may reflect the competing demands placed on facial dynamics by
561 speech articulation and emotional expression resulting in more complex
562 spatiotemporal patterns. However, while we included the facial dynamics of speech,
563 this study was not optimised to account for the vocal or prosodic dynamics of speech.
564 Given that social signalling is inherently multimodal—with tone of voice, vocal prosody,
565 and facial expressions capable of conveying emotion both independently and in
566 combination— a future challenge is to understand the extent to which these modalities
567 act redundantly (conveying the same information) or synergistically (combining to
568 produce qualitatively richer meaning) for emotion signalling (1, 9, 46).
569
570 There are three important caveats that warrant consideration. First, including more
571 detailed descriptors of face movements (including head motion see (36)) and speech-
572 optimised facial dynamics, may require additional dimensions to capture the full
573 complexity of facial behaviour. Similarly, different emotional expressions (e.g.,
574 surprise, fear, disgust), free spoken and conversational contexts, and inclusion of
575 multimodal cues may also demand higher-dimensional representations. Second, a
576 different number of substates might be discovered using physiological approaches,
577 such as electromyography (EMG). Such approaches are sensitive to subtle motor
578 potentials and therefore may capture more substates including preparatory or
579 corrective movement impulses not apparent in FACS resolution. Lastly, our approach
580 focuses on exhaustively describing spatiotemporal features, such as curvature

patterns and entropy, that account for variance in the signal. However, it does not isolate specific features (or groups of features) nor assess their individual diagnostic value for classifying emotions. While such analysis is beyond the scope of this study, it can be easily integrated with our approach.

Despite these caveats, our results and approach contribute new insights for future research on facial expression signalling. Spatiotemporal structure and substates might be important to better understand the mechanisms underlying differences in social communication. This may provide useful insight into clinical conditions such as autism and depression particularly in exploring the potential role of atypical facial expression production in contributing to social interaction difficulties (47, 48). These findings also have practical applications in fields such as social robotics and human-computer interaction. A significant challenge in these fields relates to endowing social agents (robots, digital agents) with convincing and adaptable repertoires of social behaviour, including facial signalling (49). Spatiotemporal components and substates underlying facial expression production could provide actionable targets to enhance the design of flexible realistic and expressive agents without the need to painstakingly craft each signal feature into the architecture of these systems.

Conclusion

In conclusion, we revealed a low-dimensional spatiotemporal structure with distinct substates that underpin emotional signalling in in *Expression only* and *Emotive speech* facial dynamics. We provide a framework to advance the development of novel theoretical and methodological accounts of the structure and function of expressive facial dynamics. This framework can be extended to other modalities and has practical

implications for understanding emotional signalling and perception in human and machine systems.

Materials and Methods

Participants

43 healthy volunteers from the university community (24 female; mean age = 26), took part in a study for course credit or monetary incentive. All participants gave informed consent, and all procedures of the study were approved by the institutional ethics committee. The dataset in this study was collected as part of a larger project investigating facial expression production kinematics and perception – see (11). However, all the variables and analyses reported here are novel and specific to this study and not discussed in any previous papers.

Experimental procedure

In two separate sessions, across two days, participants' facial movements were captured during a facial expression production task. Participants were asked to signal three emotional expressions: anger, happiness, and sadness in two counterbalanced conditions. An *Expression only* condition where participants simply produced facial movements to signal each expression, and an *Emotive speech* condition where they produce a fixed spoken sentence in emotional fashion (e.g., “My name is John and I’m a scientist”). The conditions are motivated by the fact that in social interaction, facial expressions can be signalled both in isolation and embedded within *Emotive speech*, meaning that facial movements might signal information relevant to emotion and speech simultaneously, yet how this is achieved in unclear. We focused on those expressions because they represent basic emotion signals which are often expressed

631 in social interaction and communication (50). While we focus on three emotions only,
632 the inclusion of a wider range of participants, speech and non-speech conditions, and
633 data across multiple days provides richer representations of facial dynamics and
634 variability compared to previous work in this area, which is often restricted to a few
635 signallers, emotions, and conditions.
636 Facial movements were recorded with participants seated with their head at the centre
637 of a polystyrene frame to ensure central positioning towards the camera, positioned 1
638 meter from a tripod-mounted video camcorder, sampling at 60Hz f/s (Sony Handicap
639 HDR-CX240E). The start and end points were signalled using an audible beep, which
640 allowed us to align the recordings of facial expressions. See the schematic illustration
641 in Fig 7.
642
643 **Analytical approach**
644 *Data processing.* Facial expression videos were processed using OpenFace (51), an
645 open-source suite for automated face-tracking, landmarking and facial action unit
646 detection based on the facial action coding system (FACS) (2). We focused on Action
647 Units (AUs), which code for the observable presence and intensity of basic face
648 actions (e.g., blinking, mouth widening, eyebrow-raising) that can be used to produce
649 almost any facial expression. We estimated the temporal activation weights for 18
650 AUs. Only videos with a tracking accuracy of over 90% and with participant data across
651 two sessions were used, resulting in a total of 532 videos with approximately 240 time
652 points each. Based on visual inspection of the movement profiles, we applied a moving
653 average (window = 3 frames) to smooth out the slight jitter inherent in face tracking
654 and AU estimation, while preserving the underlying trends and dynamic range. To
655 ensure an equal temporal window for spatiotemporal analyses, AU timeseries were

reduced into 100-time bins for each video. Binning was preferred over more computationally expensive alternatives such as dynamic time-warping (DTW) for practicality since the overall timeseries duration was very similar across videos, and visual comparisons between binning and DTW tests on our data showed negligible differences.

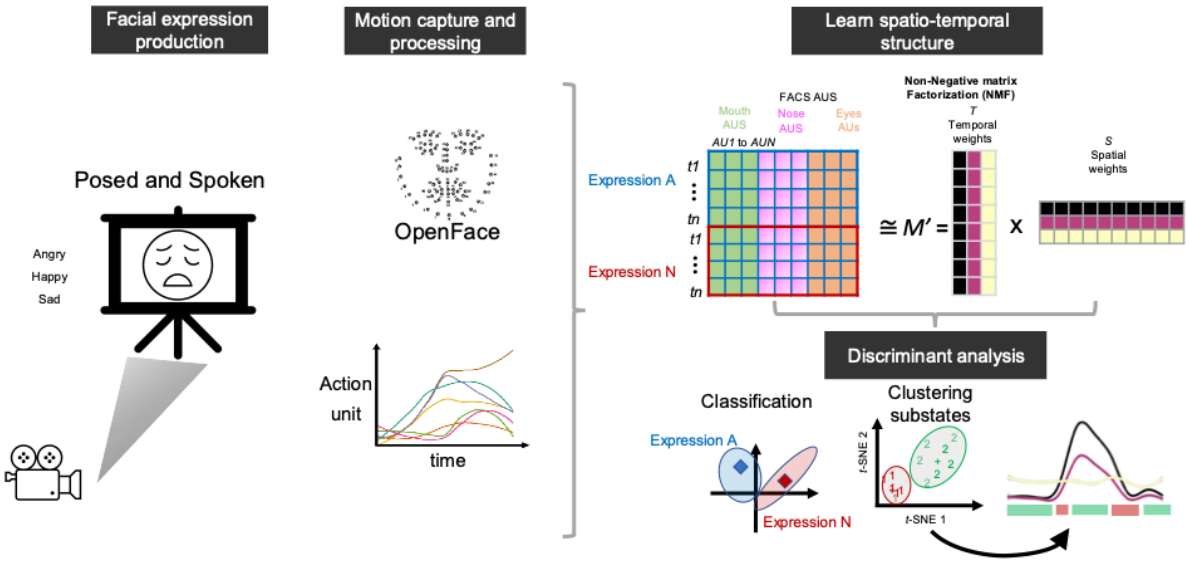

Figure 7. Schematic illustration of methodological and analytical approach. Left panel illustrates the facial expression production task, video capture and OpenFace processing for landmark and action unit (AU) estimation. The right panel summarises the procedure for the concatenation of AU timeseries data and the application of Non-Negative Matrix Factorization (NMF) to learn an underlying low-dimensional spatiotemporal representation of the data. The accuracy and effectiveness of this lower-dimensional representation were validated through a series of sensitivity and classification analysis (lower right panel). Time series clustering was used to identify distinct substates or 'phases' in facial movement components.

Spatiotemporal analysis

Timeseries of AU weights for the full facial expression production task were analysed using Non-Negative Matrix Factorization (NMF) (52). NMF is a multivariate dimensionality reduction method that decomposes a high-dimensional dataset into two non-negative matrices that form a part-based representation of the data (i.e., the original data is achieved by combining/adding components). The non-negativity

constraint is particularly well suited for representing movement and AU data, which is typically positive, and the two-matrix part-based decomposition can intuitively represent spatial vs time-varying aspects in the data – see Fig 7 (see also: (18, 34, 52)). The NMF takes the concatenated AU timeseries as input, denoted by a matrix ***M,*** which is a product of two low-dimensional matrices (***T*** and ***S***), ***T*** represents the temporal weights (basis values for time bins), and ***S*** represents the spatial weights (face action activation weights). The product of ***T*** *and* ***S*** matrices denoted by ***M'*** represents an approximation of the full original matrix such that ***M~M' = T * S.*** The algorithm works by initialising the values for ***T*** and ***S*** (e.g., with random values) and using an iterative multiplicative update rule to sequentially update *T* and *S* by a factor that minimises the approximation error (i.e., residuals). While the iterative nature of NMF means that the solution is non-deterministic (i.e., you can learn multiple equally effective part-based representations), these update rules have been shown to guarantee convergence (i.e., stability of the cost function) after a relatively low number of iterations (53). The optimal number of spatiotemporal components (*k*) was assessed computationally by fitting a range of values (*k* = 2 to 6) and choosing a *k* that optimised various model fit metrics (e.g., residual sum of squares, silhouette for clustering cohesion, cophenetic coefficient, see below). We also evaluated the final solution against NMF results on a permutated dataset, using block-based time shuffling. This resulted in an appropriate null dataset that destroyed the inherent spatiotemporal pattern in the data but retained some temporal ordering. We then compared fit metrics between original and permutated NMF (see also *Classification and clustering analysis*). We used the NMF decomposition functions in the NMF package for R (54). To visualise our NMF results in an intuitive manner, we trained a custom model to predict aligned Facial Landmark displacements (after correcting for rotation,

translation and looming) from AUs, based on original data using partial least squares regression. Note however that the AU to Landmark face model is not a perfect one to one map between NMF results and visualisation, given the many-to-one relationship between facial landmarks and AUs.

Multivariate timeseries classification

We used a supervised classification approach to validate the diagnostic value of spatiotemporal components for emotion expression. This involved five steps:

(1) Data splitting: We began by splitting the original pre-NMF data into training (80%) and testing (20%) sets, ensuring that each set contained complete AU time series for each video. This split was performed before any further analyses, and all subsequent steps maintained this separation ensuring further processing and model training was only done on the training set, and then applying results to the test set (55, 56). Hence any mention of “training”, always refers to operations on the respective training set.

(**2) NMF re-training and projection:** We then re-trained the NMF model using only the training set data and projected the test set onto the trained NMF model space to ensure consistency of components for classification.

(**3) Feature extraction:** Because standard classification approaches do not account for inherent spatiotemporal structure of the NMF components, we opted for an approach that represents the spatiotemporal structure via interpretable theoretic time series features (e.g., curvature, entropy, predictability, peak, trend) (57) consistent with features suggested to capture diagnostic information for classification of facial expressions and body movements (8, 11, 58, 59). Feature extraction was done separately for train and test sets. Then, a Principal Component Analysis (PCA) was used to reduce the timeseries features that explained the majority of the variance in

729 the train set while retaining a number of dimensions equal to the number of NMF
730 components for consistency. The test set features were then transformed using the
731 PCA model fitted on the training set features to be used in the next step. The
732 component scores were used as inputs for classification analysis in the next step.
733 **(4) Classification training and testing.** We the trained a classifier on spatiotemporal
734 component features derived from step 3 and applied to predict emotion expression in
735 a held-out test set. Specifically we used a Random Forests (RF) Classifier - a
736 supervised learning method that learns a discriminant model between a set of input
737 variables (e.g., spatiotemporal components) and a set of target classes (expressions).
738 Note, that we achieved similar results with different methods (e.g., Support Vector
739 Machine, Linear Discriminant Analysis), thus our results are not a function of the
740 analytical approach. The Random Forests Classifier has the advantage of being
741 relatively more robust to overfitting and able to capture non-linear relationships and
742 therefore was preferred. We used cross-validation during training, where multiple
743 training sets are created, and the data not included in the test set is used to test the
744 accuracy and tune the model parameters. Reliable classification on the test set would
745 indicate that the underlying spatiotemporal components (identified by the NMF
746 analysis) carry diagnostic information for differentiating expressions. We derived
747 significance tests for our classification results using binomial tests to determine
748 whether the model's accuracy on the test set was significantly greater than what would
749 be expected by random guessing (by comparing it to a no information rate). These
750 binomial p-values were compared with permutation-based p-values, which were
751 generated by comparing the observed accuracy to the distribution of accuracies
752 obtained from shuffling the labels (expressions) in the test set (N = 1000). The
753 negligible differences between the binomial and permutation-based p-values for our

754 dataset validated the robustness of our results, so we primarily report the binomial p-
755 values. Additionally, we used Fisher's Exact Test to assess the statistical significance
756 of the confusion matrix, by comparing pairwise expression differentiation with
757 adjustments for multiple comparisons. The null hypothesis is that there is no
758 association between rows and columns, or in other words differences in cell values
759 (frequency) are due to chance. A low p-value in Fisher's exact test indicates
760 statistically significant differences between pairs, meaning the pattern in the confusion
761 matrix is unlikely to have occurred by chance.
762 While a handful of previous studies previous work has employed elements of our
763 proposed pipeline such as the dimensionality reduction of facial movement data
764 across entire datasets, our approach is unique in that it explicitly builds in the
765 generalisability and robustness of the spatiotemporal patterns we uncover.
766 Furthermore, the spatiotemporal classification pipeline enables us to directly link the
767 inherent dynamics (represented through the extracted features) to their diagnostic
768 value in differentiating emotions.
769
770 *Spatiotemporal clustering for substates analyses*
771 We investigated whether reliable substates could be extracted from the spatiotemporal
772 patterns of *Expression only* and *Emotive speech* facial movements. We considered
773 substates as local intervals where the components exhibit salient signatures that are
774 consistent across expression timeseries. This can be through of as a segmentation of
775 a multivariate timeseries into finer temporal segments with arbitrary lengths. To do this
776 we first derived speed and displacement for spatiotemporal components, as these
777 features are sensitive to segment characteristic of other body movements (e.g., swing
778 and stance signatures in walking, fixation and saccades in eye movements) (19, 28,

29, 60). Second, we applied spatiotemporal clustering using the *dtwclust* package in R. This was motivated by the fact that, unlike other movements with well-established kinematic profiles, we have no *a priori* knowledge regarding the characteristics of facial expression substates. Given the computational complexity of DTW, we tested cluster sizes on a smaller subset of *Expression only and Emotive speech* data. We used visualisations which suggested (Fig 5) two to three interpretable segments described transition and stability phases in spatiotemporal components of facial expressions in line with the hypothesised patterns of transition, sustainment and relaxation of AUS. To describe the pattern of substates and whether they differentiate emotion and conditions, we computed two metrics: (1) speed within the substate and (2) entropy; both which have been shown to capture diagnostic information about facial expression movement patterns (8, 11). Specifically given that speed contributes to the clustering, statistically comparing speed give insights into how the clusters were generated (see Fig 5). Entropy, an information-theoretic measure of complexity, is directly related the dynamics of state transitions (or sequences) within a system (37). In this context, entropy provides a suitable summary of the dynamics of substates in a single statistic, describing how complex or unpredictable the sequence of these substates is (e.g., how much information about the current substate can predict the next substate). This statistic aimed at quantifying, for example, the apparent differences in the temporal patterns of substates between expressions and conditions (see Fig 5). We computed entropy using the transition entropy function (for state change entropy) from the *GrpString* package for R (61) but for simplicity we refer it as entropy. To improve interpretability, entropy was normalised by dividing it by the maximum theoretical entropy for a system with the same number of substates.

Both speed and entropy measures were computed for each individual expression and analysed using a mixed-effects model to test the effect of emotion (happy, angry, sad), condition (*Expression only*, *Emotive speech*), and substate (relaxed, transition, sustain) as fixed effects and subject (face id) as a random effect. Entropy was log transformed to correct for non-normal residuals. Substate metrics were analysed using linear mixed models as opposed to the classification pipeline, due to the more complex and nested structure of the data which also makes classification overly complex to structure and present. Linear mixed models were fitted using the lme4 package for R (62).

Data availability statement

Supporting data is available in the Open Science Framework (OSF): https://osf.io/3nzk7/?view_only=db7d9467423f42b794385e549b6aa96a

All code for pre-processing and models is available at https://github.com/hcuve/spat-tmp-emo-faces

Acknowledgments

We would like to thank Professor Rachael Jack for her invaluable insights during the initial discussions on facial signalling dynamics and dimensionality reduction.

We would also like to thank Katherine Lane and Tahniyyat Bokhari, for their assistance with data collection for a portion of the study.

Author Contributions (CRediT)

HCC, Conceptualisation, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Validation, Visualisation, Writing – Original Draft Preparation, Writing – Review & Editing

SSC, Methodology, Validation, Writing – Review & Editing, Resources

JC, Conceptualization, Funding Acquisition, Methodology, Writing – Review & Editing, Resources

Funding Statement

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under ERC-2017-STG Grant Agreement No 757583. HCJC was funded by an Experimental Psychology Society Postdoctoral Fellowship

References

1. R. E. Jack, P. G. Schyns, The human face as a dynamic tool for social communication. *Curr. Biol.* **25**, R621-34 (2015).

2. P. Ekman, W. V. Friesen, Facial action coding system. *Environmental Psychology & Nonverbal* (1978).

3. E. G. Krumhuber, L. I. Skora, H. C. H. Hill, K. Lander, The role of facial movements in emotion recognition. *Nat. Rev. Psychol.* (2023) https:/doi.org/10.1038/s44159-023-00172-1.

4. K. R. Scherer, H. Ellgring, Multimodal expression of emotion: affect programs or componential appraisal patterns? *Emotion* **7**, 158–171 (2007).

5. A. A. Ghazanfar, D. Y. Takahashi, Facial expressions and the evolution of the speech rhythm. *J. Cogn. Neurosci.* **26**, 1196–1207 (2014).

6. C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, A. A. Ghazanfar, The natural statistics of audiovisual speech. *PLoS Comput. Biol.* **5**, e1000436 (2009).

7. R. E. Jack, P. G. Schyns, Toward a social psychophysics of face communication. *Annu. Rev. Psychol.* **68**, 269–297 (2017).

8. R. E. Jack, O. G. B. Garrod, P. G. Schyns, Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Curr. Biol.* **24**, 187–192 (2014).

9. M. Liu, *et al.*, Facial expressions elicit multiplexed perceptions of emotion categories and dimensions. *Curr. Biol.* **32**, 200-209.e6 (2022).

10. D. Gill, O. G. B. Garrod, R. E. Jack, P. G. Schyns, Facial movements strategically camouflage involuntary social signals of face morphology. *Psychol. Sci.* **25**, 1079–1086 (2014).

11. S. Sowden, B. A. Schuster, C. T. Keating, D. S. Fraser, J. L. Cook, The role of movement kinematics in facial emotion expression production and recognition. *Emotion* **21**, 1041–1061 (2021).

12. A. M. Burrows, The facial expression musculature in primates and its evolutionary significance. *Bioessays* **30**, 212–225 (2008).

13. C. C. Sherwood, Comparative anatomy of the facial motor nucleus in mammals, with an analysis of neuron numbers in primates. *Anat. Rec. A Discov. Mol. Cell. Evol. Biol.* **287**, 1067–1079 (2005).

14. E. Todorov, Optimality principles in sensorimotor control. *Nat. Neurosci.* **7**, 907–915 (2004).

15. R. Pang, B. J. Lansdell, A. L. Fairhall, Dimensionality reduction in neuroscience. *Curr. Biol.* **26**, R656-60 (2016).

16. M. L. Latash, J. P. Scholz, G. Schöner, Toward a new theory of motor synergies. *Motor Control* **11**, 276–308 (2007).

17. E. Chiovetto, A. Salatiello, A. d'Avella, M. A. Giese, Toward a unifying framework for the modeling and identification of motor primitives. *Front. Comput. Neurosci.* **16**, 926345 (2022).

18. E. Chiovetto, C. Curio, D. Endres, M. Giese, Perceptual integration of kinematic components in the recognition of emotional facial expressions. *J. Vis.* **18**, 13 (2018).

19. O. Labaune, T. Deroche, C. Teulier, B. Berret, Vigor of reaching, walking, and gazing movements: on the consistency of interindividual differences. *J. Neurophysiol.* **123**, 234–242 (2020).

20. Y. P. Ivanenko, R. E. Poppele, F. Lacquaniti, Five basic muscle activation patterns account for muscle activity during human locomotion. *J Physiol (Lond)* **556**, 267–282 (2004).

21. E. Chiovetto, B. Berret, T. Pozzo, Tri-dimensional and triphasic muscle organization of whole-body pointing movements. *Neuroscience* **170**, 1223–1238 (2010).

22. A. Zangrossi, G. Cona, M. Celli, M. Zorzi, M. Corbetta, Visual exploration dynamics are low-dimensional and driven by intrinsic factors. *Commun. Biol.* **4**, 1100 (2021).

23. V. Ramanarayanan, L. Goldstein, S. S. Narayanan, Spatio-temporal articulatory movement primitives during speech production: extraction, interpretation, and validation. *J. Acoust. Soc. Am.* **134**, 1378–1394 (2013).

24. A. Barliya, L. Omlor, M. A. Giese, A. Berthoz, T. Flash, Expression of emotion in the kinematics of locomotion. *Exp. Brain Res.* **225**, 159–176 (2013).

25. J. M. Montepare, S. B. Goldstein, A. Clausen, The identification of emotions from gait information. *J. Nonverbal Behav.* **11**, 33–42 (1987).

26. E. S. Dalmaijer, A. Lee, R. Leiter, Z. Brown, T. Armstrong, Forever yuck: Oculomotor avoidance of disgusting stimuli resists habituation. *J. Exp. Psychol. Gen.* **150**, 1598–1611 (2021).

27. E. L. Mackevicius, *et al.*, Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. *eLife* **8** (2019).

28. M. Rucci, M. Poletti, Control and Functions of Fixational Eye Movements. *Annu. Rev. Vis. Sci.* **1**, 499–518 (2015).

29. F. Torricelli, *et al.*, Motor invariants in action execution and perception. *Phys. Life Rev.* **44**, 13–47 (2023).

30. E. G. Krumhuber, K. R. Scherer, Affect bursts: dynamic patterns of facial expression. *Emotion* **11**, 825–841 (2011).

31. K. L. Schmidt, J. F. Cohn, Y. Tian, Signal characteristics of spontaneous facial expressions: automatic movement in solitary and social smiles. *Biol. Psychol.* **65**, 49–66 (2003).

32. S. Aghabozorgi, A. Seyed Shirkhorshidi, T. Ying Wah, Time-series clustering – A decade review. *Inf. Syst.* **53**, 16–38 (2015).

33. C. C. Aggarwal, “An introduction to cluster analysis” in *Data Clustering: Algorithms and Applications*, C. C. Aggarwal, C. K. Reddy, Eds. (Chapman and Hall/CRC, 2018), pp. 1–28.

34. I. Delis, *et al.*, Space-by-time manifold representation of dynamic facial expressions for emotion categorization. *J. Vis.* **16**, 14 (2016).

35. S. Namba, K. Nakamura, K. Watanabe, The spatio-temporal features of perceived-as-genuine and deliberate expressions. *PLoS ONE* **17**, e0271047 (2022).

36. D. M. Watson, B. B. Brown, A. Johnston, A data-driven characterisation of natural facial expressions when giving good and bad news. *PLoS Comput. Biol.* **16**, e1008335 (2020).

37. C. E. Shannon, W. Weaver, *The mathematical theory of communication* (books.google.com, 1998).

38. T. Le Mau, *et al.*, Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. *Nat. Commun.* **12**, 5037 (2021).

39. B. Parkinson, “Social Functions” in *Heart to Heart: How Your Emotions Affect Other People*, (Cambridge University Press, 2019), pp. 215–240.

40. A. W. Siegman, S. Feldstein, Eds., *Multichannel integrations of nonverbal behavior* (Psychology Press, 2014) https:/doi.org/10.4324/9781315802305.

41. M. S. Gobel, H. S. Kim, D. C. Richardson, The dual function of social gaze. *Cognition* **136**, 359–364 (2015).

42. D. M. Wolpert, K. Doya, M. Kawato, A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **358**, 593–602 (2003).

43. S. M. Bell, M. D. Ainsworth, Infant crying and maternal responsiveness. *Child Dev.* **43**, 1171–1190 (1972).

44. C. Rowe, Receiver psychology and the evolution of multicomponent signals. *Anim. Behav.* **58**, 921–931 (1999).

45. G. B. Duchenne de Boulogne, *The mechanism of human facial expression*, R. A. Cuthbertson, Ed. (Cambridge University Press, 1990) https:/doi.org/10.1017/CBO9780511752841.

46. E. A. Hebets, *et al.*, A systems approach to animal communication. *Proc. Biol. Sci.* **283**, 20152889 (2016).

47. R. Brewer, *et al.*, Can neurotypical individuals read autistic facial expressions? atypical production of emotional facial expressions in autism spectrum disorders. *Autism Res.* **9**, 262–271 (2016).

48. A. Haque, L. Fei-Fei, M. Guo, A. S. Miner, Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions. *Machine Learning for Health (ML4H) Workshop at NeurIPS 2018, Montréal, Canada* (2018).

49. N. Rawal, R. M. Stock-Homburg, Facial emotion expressions in human–robot interaction: A survey. *Int. J. Soc. Robot.* **14**, 1583–1604 (2022).

50. M. G. Calvo, A. Gutiérrez-García, A. Fernández-Martín, L. Nummenmaa, Recognition of Facial Expressions of Emotion is Related to their Frequency in Everyday Life. *J. Nonverbal Behav.* **38**, 549–567 (2014).

51. T. Baltrušaitis, P. Robinson, Openface: an open source facial behavior analysis toolkit. *2016 IEEE Winter* (2016).

52. D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

53. X. Fu, K. Huang, N. D. Sidiropoulos, W.-K. Ma, Nonnegative matrix factorization for signal and data analytics: identifiability, algorithms, and applications. *IEEE Signal Process. Mag.* **36**, 59–80 (2019).

54. R. Gaujoux, C. Seoighe, A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).

55. N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, C. I. Baker, Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12**, 535–540 (2009).

56. G. Varoquaux, O. Colliot, “Evaluating machine learning models and their diagnostic value” in *Machine Learning for Brain Disorders*, O. Colliot, Ed. (Humana, 2023) https:/doi.org/10.1007/978-1-0716-3195-9_20.

57. F. J. Baldán, J. M. Benítez, Multivariate times series classification through an interpretable representation. *Inf Sci (Ny)* **569**, 596–614 (2021).

58. K. Dobs, *et al.*, Quantifying human sensitivity to spatio-temporal information in dynamic faces. *Vision Res.* **100**, 78–87 (2014).

59. N. Furl, *et al.*, Face space representations of movement. *Neuroimage* **212**, 116676 (2020).

60. J. G. V. Miranda, *et al.*, Complex Upper-Limb Movements Are Generated by Combining Motor Primitives that Scale with the Movement Size. *Sci. Rep.* **8**, 12918 (2018).

61. I. Hooge, G. Camps, Scan path entropy and arrow plots: capturing scanning behavior of multiple observers. *Front. Psychol.* **4**, 996 (2013).

62. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).