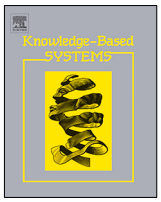
Knowledge-Based Systems 276 (2023) 110759

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Agreement and disagreement between major emotion recognition systems

Carl Vogel *, Khurshid Ahmad

School of Computer Science and Statistics, Trinity College Dublin, the University of Dublin, Ireland

ARTICLE INFO

Article history:
Received 18 April 2023
Received in revised form 9 June 2023
Accepted 25 June 2023
Available online 7 July 2023

Keywords:
Emotion processing
Emotion recognition systems
Multi-modal communication data

ABSTRACT

The evaluation of systems that claim to recognize emotions expressed by human beings is a contested and complex task: The early pioneers in this field gave the impression that these systems will eventually recognize a flash of anger, suppressed glee/happiness, momentary disgust or contempt, lurking fear, or sadness in someone's face or voice (Picard and Klein, 2002; Schuller et al., 2011). Emotion recognition systems are trained on 'labelled' databases — collection of video/audio recording comprising images and voices of humans enacting one emotional state. Machine learning programmes then regress the pixel distributions or wave forms against the labels. The system is then said to have learnt how to recognize and interpret human emotions and rated using information science metrics. These systems are adopted by the world at large for applications ranging from autistic spectrum communications to teaching and learning, and onwards to covert surveillance. The training databases depend upon human emotions recorded in ideal conditions — faces looking at the camera and centrally located, voices articulated through noise-cancelling microphones. Yet there are reports that the posed training data set, that is racially-skewed and gender unbalanced, does not prepare these systems to cope with data-in-the-wild and that expression-unrelated variations (like illumination, head pose, and identity bias (Li and Deng, 2020)) can impact their performance as well. Deployments of these systems tend to adopt one or other and apply it to data collected outside laboratory conditions and use the resulting classifications in subsequent processing. We have devised a testing method that helps to quantify the similarities and differences of facial emotion recognition systems (FER) and speech emotion recognition systems (SER). We report on the development of a data base comprising videos and sound track of 64 politicians and 7 government spokespersons (25 F, 46 M; 34 White Europeans, 19 East Asians, and 18 South Asians), ranging in age from 32–85 years, and each of the 71 has on average three 180 s videos; a total of 16.66 h of data. We have compared the performance of two FERs (Emotient and Affectiva) and two SERs (OpenSmile and Vokaturi) on our data by analysing emotions reported by these systems on a frame-by-frame basis. We have analysed the directly observable head movements, and the indirectly observable muscle movement parts of the face and for the muscle movements in the vocal tract. There was marked disagreement in emotions recognized, and the differences were exacerbated more women than for men, and more for South and East Asians than for White Europeans. Levels of agreement and disagreement on both high-level (i.e. emotion labels) and lower-level features (e.g. Euler angles of head movement) are shown. We show that inter-system disagreement may also be used as an effective response variable in reasoning about data features that influence disagreement. We argue that reliability of subsequent processing in approaches that adopt these systems may be enhanced by restricting action to cases where systems agree within a given tolerance level. This paper may be considered as a foray into the greater debate about the so-called algorithmic (un)fairness and data bias in the development and deployment of machine learning systems of which FERs and SERs are a good exemplar.

1. Introduction

The claims and the promise of automatic facial emotion recognition (FER), and speech emotion recognition (SER), are irresistible both intellectually and commercially.[1] One need not look

* Corresponding author.
E-mail address: vogel@cs.tcd.ie (C. Vogel).

[1] See, for example, [1] and [2].

https://doi.org/10.1016/j.knosys.2023.110759

far to identify research that has adopted an SER or FER system as a component in larger systems or behavioural analysis methods.[2] In general terms, we can have a fair idea of a familiar or unfamiliar person's emotional state irrespective of their intrinsic age, gender, race/colour, luminosity, occlusion, and orientation. Acting in context, humans behave as if they are relatively good at judging emotional states of others within the contexts on the basis of non-verbal behaviours, even without access to linguistic content expressed. Of course, the basis of many miscommunications and conflicts is incorrect perception of the emotional states of others in shared contexts, and human judgements of non-shared contexts, such as in analysing video, are less stable (cf. [3]). This we do just by looking at the face and mentally computing the reflection and absorption of light on the facial skin and hair, focussing on key small areas on the skin to detect the presence of a face and the changes in reflection/absorption over time providing clues about the emotional state of the observed person. This is somewhat true of hearing a stranger who speaks another language to ours, and getting a glimpse into their emotional state just by detecting and noticing the changes in the prosody, vocal energy, and voice quality in their speech. This we humans do without even looking at the speaker in the face; one can even make a reasonable guess of their age, gender, personality, forcefulness, origins, without any facial encounter. In the case of both FER and SER, we have to recognize the existence of a face, or of a voice, focus on key areas of the face, or key properties of the voice, relate what we are seeing now has to what we have seen/heard, and then to predicate what we might see/hear. Finally, we assign a label to the emotion on the face or in the voice. The emotional model which has dominated the emotion recognition literature suggests that the movement of facial muscles, and muscles use to articulate voice, are what we human focus on. These proponents of these models are equally emphatic that humans have 5–7 seven 'basic' emotions that relate to muscle movements in an idiosyncratic fashion. The widely used FER and SER systems are based on the muscle dynamics and basic emotion theory. In this paper, while we describe some prior work on emotion recognition that attends to linguistic content, text emotion recognition, our focus is on the systems that assess acoustic properties of speech and visual properties of the face and head, but without reference to text emotion classification.

Speech producing muscles, around 100 or more in number, contract fast and at variable speeds, are fatigue resistant, and have different biomechanical and histological properties [4]. Speech producing muscles include craniofacial muscles — muscles controlling the lips and mouth corners, and muscles that control the larynx (simply known as the voice box) are unique in the human muscular system. The muscles act in agonist and antagonist modes — maintaining an equilibrium between a contracting muscle and a relaxing muscle. It is the collaboration and opposition of these muscles that creates speech. The physical properties of the speech (waveform) include frequency, energy, are indirect measures of the muscle dynamics. The muscles age and voice changes with aging. Gender and race are the key variables for some vocal tract properties – and these properties [5] lead to differences in the speech of men and women, and in the speech of different racial groups – causing the so-called gender and racial disparities in voice biometrics [6].

Facial muscles – around 42 in number – are subcutaneous (under the skin) and are unique in that these muscles are innervated by facial nerves and control facial expressions. Facial muscles control the three dimensional face deformation caused by the movement of 'antagonistic and synergistic muscles' causing movements in and around the face by contraction and relaxation of muscles [7]. Facial recognition and facial expression is indirectly calculated by the movement of proxies called the facial landmarks that have specific optical properties [8]. Again the skin distortion/displacement is an indirect measure of the muscle dynamic. The facial skin has its own undulations due to differences in facial bone structure and skin thickness when Caucasians and Japanese Asian faces are compared [9]. Facial anthropometric studies suggest that there are statistically significant differences in nose dimensions amongst African Americans, White Americans, Asians and Hispanics [10] and that gender is a key variable in this respect as well [10].

That the movement of facial muscles [11,12], and vocal tract muscles [13], can be successfully used as a correlate of human emotion is a very interesting example of the use of Occam's Razor. Note that facial anthropometry, a mainstay of craniofacial forensic analysis [14,15] and of plastic surgery [16–18], is the intellectual precursor of automatic face recognition. Acoustic analysis, a mainstay of voice biometric systems [5] (and are used in forensic analysis of voice [19]) is the intellectual precursor of speech recognition systems. The technological precursors of FER and SER include the automatic face recognition systems and speech recognition systems respectively.

These two modalities of non-verbal communication can reinforce emotions expressed in each other or may contradict the emotions. The analogy of twisted fibres is useful: each modality provides a fibre within a string, and which is most prominent depends on the perspective one has to the whole; further, the fibres may be twisted with differing degrees of tightness about each other, such that within a given length and from a certain perspective, one fibre may be visible in more shorter bursts, and others in fewer, longer lengths. So, too, emotive content may be expressed within entwined modalities, the possibilities for change in emotive content in any modality adding more possible variation to what is observed from any given perspective. Our intuition about the twisted pair of modalities has motivated us discuss these together in this paper.

The current ontological basis of facial emotion recognition systems is rooted in the existence of an ideal face with ideal facial features which can be detected by face and facial feature detection algorithms and matched against the contents of a corpus of acted emotions [20–22]. Much the same is true of speech emotion recognition system predicated on the existence of an ideal voice with ideal features which can be detected by voice and vocal feature detection algorithms and matched against a labelled corpus of soundtracks of actors voicing emotions [23–25]. The universality arguments of ideals and algorithm have led to a whole range of laboratory experiments conducted by intellectually gifted scientists and engineers on state of the art devices which appear to extract hitherto unquantifiable and largely uncomputable human emotions with a degree of fidelity. So far so good. But the laboratory experiments are now on the valorization juggernaut and we have commercially available facial emotion recognition systems and speech emotion recognition systems. The usage of these systems knows no bound — from the needy area of surreptitious surveillance and onto diagnostic medicine, especially autism spectrum disorder, and from marketing to evaluating educational attainment, from robots in smart manufacturing to robots in social care. There are important technological issues about the accuracy of such systems on the one hand, and societal and moral reservations about the automatic emotion recognition system on the other.

The recognition and interpretation of power of FERs and SERs is usually stated in precision/recall terms of what the training label of a video (audio track) is and what an FER/SER outputs it to be. In order to pursue this strategy one has to place great faith in the design and original annotations of the training (and testing) databases. The inter-annotator agreement is usually 'substantial'

[2] See Section 2.4, where we point out some of these.

(using Cohen's Kappa) across the various emotional states [26,27]. FERs depend on the location and activity of the so-called action units on the face of a person who is expressing one emotion. This task is carried out by trained facial action coding systems experts. Their handcrafted observations are used in training FERs to relate the changes in the activity of action units (AUs) over time (typically milliseconds). The behaviour of some is readily learnt with high precision but for others the situation is not so satisfactory [28]; automated facial coding (AFC) has had more success with 'near perfect agreement' for very specific parts of the face which whilst for others the agreement between AFC and a FAC coder varies between 'moderate to 'substantial' agreement [29]. Note that usually several action units are activated when people experience an emotion, so we are relying on a variable recognition performance per AU which, in turn, has an impact on the accurate recognition and interpretation of an image not seen by an FER.

We have taken a different approach to evaluating the FERs and SERs. Instead of comparing systems in relation to precision and recall on extant "gold-standard" datasets, or creating a corpus of videos with accompanying texts and labelling as a novel "gold-standard", we have created a corpus of a (in-the-wild) videos of politicians, their spokespersons, and CEOs, of different races, genders and ages, with the purpose of comparing SER and FER system assessments of such data. We chose our cohort as their job is to persuade us to follow a policy with charm, authority, and empathy and they use the ebb and flow of emotions as a persuasive device. For each of our cohort we have 3 or more videos.[3] We have used the videos as inputs to two different FERs (Emotient and Affectiva) and found the outputs statistically-significantly different for key emotions. The same is true of the two SERs (OpenSmile and OpenVokaturi) that were given the sound tracks as the input. These systems are selected for analysis here because each has been freely available for use and each has been widely used (as we note below), even though more advanced systems are now available through remunerated licenses. The systems we analyse are architecturally different and use different training databases for emotion recognition. The differences in the outputs are important because of the various applications that have used just the one FER (or SER) and made various claims about (individual or group) human behaviour. For if there is a slightest of doubts about the outputs of an automatic system, then the use of the outputs should pose a serious moral dilemma to the enthusiastic proponents of FER and SER. Our contrastive methodology may help in bringing forward important research advances to the public forum but with caveats that must accompany any computer system. We focus here on analysis of system behaviours on the data that we have collected rather than on "gold standard" data sets. One reason is that emotion recognition systems are likely to be deployed on data captured "in the wild" rather than with repeated testing on the original training data or other gold-standard data sets. Another reason, as we discuss later, is that there is ample reason to doubt the possibility of gold-standard emotion labels on data: acted emotion is not the same as authentically experienced emotion; third-party labels of experienced emotion of others are not reliable; time-delayed subjective labels of experienced motion are not reliable. Thus, we suggest that the best standard one could hope to achieve is multiple annotator agreement. Disagreement on emotion labels also suggests assessing agreement on measurements of more basic physical properties that contribute to emotion assessment, as we describe.

[3] Others also note issues that arise with data-in-the-wild and that expression-unrelated variations (like illumination, head pose, and identity bias [30]) can impact emotion recognition system performance as well.

The main contributions of this paper are in the identification of the extent to which the emotion classification systems agree and disagree on high-level and low-level features of audio and video processing. We demonstrate a means of using quantification of system disagreement as a response variable, in order to identify the features that interact with disagreement. The overall method of analysis we use may be adapted to similar tasks with the same systems on other data sets, other data sets, and comparable systems. The background corpus of videos curated in the context of this research and analysed here is also available to the research community (the corpus derived from content visible to the general public).

The rest of this paper is structured as follows. First, we contextualize our work with respect to a selection of extant voice and facial emotion recognition systems (Section 2). We describe some of the assumptions of emotion recognition systems (Section 2.2) and then some of the systems (Section 2.3). We note contributions to the literature in a variety of fields that depend on the emotion classifications provided by these systems (Section 2.4). We describe prior published evaluations and comparisons of these systems (Section 2.5), and then synthesize some of our own works in comparing them (Section 3). We discuss the ramifications of these comparisons (Section 4): because of the system-dependence of classifications, it is important for those seeking to rely on the classifications to be cautious in using them, perhaps seeking majority classification from more than one, or using them only on datasets with properties on which the systems can be demonstrated to agree or for which judgements of one may be algorithmically transformed into judgements of others. We also show that system agreement and disagreement may be used as a variable to illuminate the study of other variables represented in data to which the systems are applied. We conclude (Section 5) by noting limitations of the current state of our analysis and our overall programme while highlighting our view of the relevance of this work. Our work is a contribution towards identifying the parameters of trusted use of automatic emotion recognition systems. This paper provides a synthetic view of our earlier work [31–33], and builds on that prior research with additional analyses — in particular, we show (Section 4) the value of reasoning with system differences as a response variable and understanding features in the data that influence system differences.

A Note on the Data Used We have been working on the expression of emotions by authority figures before, during and after a natural disaster [34] since 2014. Over the years we have been increasing our video corpus every year with help from our undergraduate and postgraduate students then. The authors of this paper have supervised the collection and developed methods of analysis. We will like to express our gratitude to them; some of them have been joint authors on papers with us [31–33].

2. Related work

In this section, we describe systems for recognition of vocal and facial expressions of emotions. We note research that depends on the emotion classifications provided by these systems for other purposes. We also describe works additional to our own that have provided evaluations and comparisons of these systems.

2.1. Linguistic emotion recognition systems

First, it is necessary to acknowledge a large body of research that analyzes the emotive content of linguistic expressions as recorded in texts. Naturally, this body have work is anchored in words and wordlists, such as provided by the General Inquirer [35] and Linguistic Inquiry and Word Count (LIWC) [36]. Over

the years, these methods have been deployed widely [37–39] and cross compared [40]. More recent works take advantage of more rich representations of word meaning than available in classified word lists, through distributional semantic representations and resulting representational similarity, and richer representations of relevant contexts, including cross-speaker relations in natural language dialogue [41,42]. However, as indicated above, our focus in this work is on systems that analyse the acoustic properties of speech and other vocalizations, rather than the linguistic content, and also systems that analyse the visible behaviours of the face and head in the course of classifying emotive expression.

2.2. Precursors of non-verbal emotion recognition systems

The precursors of FER include the face recognition systems and for the SER it is the speech recognition systems. These recognition systems are trained, using machine learning algorithms, to recognize a human face and a human voice. The US National Institute for Standards and Testing has evaluated a number of facial recognition systems in the last three years and have found that these systems are based on algorithms and training regimens that may produce false positives/negatives: "Despite all of the advances in algorithm design, facial recognition algorithms have several known cultural problems in the basic identification of various faces from different races regardless of the country in which the algorithm is developed" [43]. The bias comes from using training databases that may have a larger number of faces for one racial group than others for example. The biases may have roots in how we dealt with the world in early infancy: Infants see more people of their own skin tone and physiognomy leads to an 'early visual preference and recognition advantage for the familiar race group' that persists in later life [44]. As the infants grow older social identity, prejudice, status and power leads to more bias towards own kind [45]. One key use of automatic facial recognition in policing is in limbo as the technology is perceived to be racially biased [46].

Facial skin optics is seldom discussed in the FER and face detection and analysis literature. The detection of a face using a ordinary light source involves the optical properties of the skin and this in turn determines how much light is reflected by the skin surface, epidermis and dermis, and how much light is absorbed [47]. Recent experimental studies have shown that different areas of the face – forehead, eyes, cheeks, nose and chin – have different reflectivity and absorption which varies according to skin tone or colour, gender and age, and equally depends on the RGB components [48]. The skin thickness equally varies across the face: the thickest dermis was on the lower nasal sidewall and thinnest is on the upper medial eyelid, and thickest epidermis is on the upper lip [49]. The thickness of the nasal soft tissue envelope varies according to racial categories [50]. The thickness will determine the reflectivity to a certain extent and will have an impact on face detection and recognition.

The automatic voice recognition has an equal quantum of persistent criticism regarding racial bias over the last 15 years or so [51,52]; this asymmetric effect has been attributed to the skewed nature of training database, towards one racial group, used to train these systems. Furthermore, there are studies revealing gender and dialect bias in such systems as well [6,53,54] plus well known speech recognition systems are not good in recognizing dysphonic voices [55].

Facial emotion recognition systems comprise a pipeline of systems where the video input, preferably comprising one person's face preferably, is processed and the emotion content of the video is quantified as a time series. The first two crucial systems in the pipeline are face detection and facial feature detection systems. Face detection is carried out using a variant of the Viola–Jones rigid object detection algorithm that is suitably adapted for use by neural network techniques for moving objects. Facial feature identification involves finding the so-called facial landmarks around the detected face. These strengths and weaknesses of these two systems in the pipeline are based on intuitive or analytical models of where a face is and what the features are? The other important parts of the pipeline include a system that tracks the dynamics of the landmarks on the face, and a system that matches the dynamics with an emotion labelled set of vectors. Each of these systems is based on intuitive and analytical models of the dynamics and that of the matching strategies. It is remarkable that with all the intricacies and assumptions underlying of the models that facial emotion recognition systems do as well in recognizing emotions at all.

Facial emotion systems depend equally critically on the availability of labelled data bases of humans are acting out to be in a given emotional state over a short period of time. Experts/systems designers select a corpus of videos of people who have been instructed to show a set of emotions, and then populate their databases on the attributes and values of the facial feature specified intuitively or analytically by the designers. These databases are used to train facial emotion recognition systems. These training videos, however, are not designed to cope with assertion that 'every face is different' and that every face 'reflects (sic) […] something about unique about' a human being – aspects of their 'heritage – including race, ethnicity, culture, geography'. The other confounding factor about a human face is the age, gender, and various (unique) means of self-expression. It has been suggested that the key attributes that might help create such a diverse and representative database of faces expressing emotions, should include craniofacial distances, areas, and ratios, facial symmetry and contrasts, skin colour, age, gender, and pose [56].

Surveys of architectures of FER and SER show a variety detection and analysis systems used that are based on selected facial and speech features. There is a large choice of training databases for the classification of changes in facial and speech features and the data bases have been used to train neural networks of different configurations (CNN, Deep Neural Nets, Recurrent nets, LSTM) (See [57] for SER architectures/databases/neural nets, and [58] for FER). There is a wide variation here and any two FER (SER) systems are different and yet if the two systems provide the same emotion expression output for the same input, then we are on fairly safe intellectual and technological grounds. If this hypothesis is not confirmed then we have to think about how to standardize the use of technology and training data. This is the burden of argument in this paper.

2.3. Non-verbal emotion classification systems

The Darwinian notion underpinning emotion recognition research, propagated successfully in both intellectual and commercial sense, relates to the hypothesis that 'a person's emotional state can be readily inferred from his or her facial movements, typically called emotional expressions or facial expressions'. The notion is compatible with the view of James [59] that bodily responses to stimuli (physical or conceptual) *are* the emotions. The Darwinian idea has substantial implications for all walks of life, including political and legal, commercial, safety and security, education, health, and economics/finance. One can see that these systems based on basic emotion theory, that there are six or seven basic emotions, have achieved commercial success and are being deployed in the 'real world', and have led to the publication of a number of papers in learned journals. This hypothesis is used as a theoretical specification for both facial and speech emotion recognition system has been challenged [60,61] and by a number of scholars who have shown the six emotional states

are a part of much larger sets of emotions, including amusement, boredom, shame, and confusion to name but a few, especially when conducting research in a number of cultures [62]. The work in cross-cultural studies of human emotions shows that there may be four latent emotions that are culturally more common than the six-basic emotions [63]. What we learn from this debate between the basic emotion theorists and those who prefer a different description of emotions is that emotions are shaped by (the external features) of social context and the other is the situational context — who is expressing the emotion, what is the context, how the person who is expressing emotions is dealing with power, gender, class, and origins. Our choice of the test is guided by trying to select a variety of politicians drawn from different countries, are of different ages and gender.

It has been identified that there are culturally specific dimensions to human perception of emotion [64,65], but within-culture variation is also evident, with factors such as age [66], sex [65], autism spectrum disability [67], and major depression [68], among other factors.

Researchers have examined facial emotion software systems and found that they correlate well with independent measurements of muscles implicated in facial action unit movements [69]. However, [31, p. 203] have demonstrated that widely used systems (Emotient FACET, Affectiva AFFDEX [70] and Azure) differ significantly in the determination of the emotions Joy, Fear, Disgust, Contempt and Sadness (the null hypothesis of agreement on Surprise and Anger could not be rejected).

Other systems have been developed for re-use outside their development teams, but we focus here on openSMILE/OpenEar and OpenVokaturi among SERs and Emotient FACET (to which we generally refer as "Emotient") and Affectiva AFFDEX (to which we refer as "Affectiva") and to a limited extent, Azure, among FERs. Both systems have been available relatively recently from the company iMotions – iMotions make the two systems 'available' through their system – using the software as service model. Both AFFDEX and Emotient exist in their own right but are now not available to the public directly. Sometimes the literature that we review refers to a system as "iMotions" — we have tried to use context to clarify which FER is addressed in such work.

The OpenVokaturi system is informed by the work of Boersma [71] in the Praat acoustic analysis system. The emotion labels of OpenVokaturi are applied after an initial voice processing step that identifies nine features in the acoustic signal[4]: average pitch, pitch dynamics, pitch jitter, average intensity, intensity dynamics, intensity jitter, spectral slope, spectral jitter.[5] The processing of the cues from the acoustic signals in training data analyses the measures for the cues for durations given each of the emotion labels (anger, boredom, disgust, fear or anxiety, happiness or joy, sadness and neutral), and a neural network architecture is used to identify probabilities associated with each emotion label.[6] The openSMILE System has capabilities for both video and audio processing; however, here we focus on the audio analysis, which is the original focus of the system [72,73]. A large number of features are provided, rooted in low-level acoustic feature descriptions and elaborated with a range of statistics for those features over sliding windows of sampling durations [74]. We use the "emobase" configuration from openEAR, with 998 such features, informing emotion classification using a number of learning algorithms [75].

Within facial emotion recognition, the underlying low-level classifications include face detection, head movement determination and landmark identification. The Emotient system records measurements for 20 facial action units and also head movement estimates of yaw, pitch and roll; high-level classifications provide measurements of the emotions anger, joy, fear, disgust, sadness and surprise [31,76,77]. Affectiva provides measurements supporting emotion classifications with respect to anger, disgust, fear, joy, sadness, surprise and contempt [70]. Thirteen action units and "smirk" are identified as low-level facial descriptors to inform the emotion classifications, and the Euler angles – yaw, pitch and roll – are estimated.

2.4. Deployments of non-verbal emotion classification systems

Here we note contributions to the literature which report on research that relies on emotion classification provided by extant systems using their default models. This demonstrates that researchers rely upon the judgements of these systems.

2.4.1. Speech emotion

In approaching automated classification of emotion from the voice signal, it is commonplace [78–80] to use extant systems, such as openSMILE voice signal processing toolkit, in order to extract measurements of acoustic properties of voice [72,73], and the OpenEar default feature sets and pre-trained models founded on the openSMILE work. OpenVokaturi has its origins in the Praat system for speech analytics [71].

Some research draws on the assumption that emotion classification systems provide trustworthy measurements. Ma et al. [81] use OpenVokaturi in an experiment whose premise is to evaluate the extent to which non-professional actors could perform designated emotions, as classified by the system. They conclude (p. 3), "In our study, most participants did not succeeded in mimicking all five basic emotions. However, we cannot simply claim that the SER system OpenVokaturi we used is not good enough for emotion detection. Instead, we argue that it may have failed because the participants were unable to successfully act out each emotion..". The apparent suggestion is that the system provides a higher standard of emotion classification than the human participants provide in emotion performance. The OpenVokaturi system has been adopted in work on music composition to evaluate the quality of synthetic voice emotion [82]. Ortloff et al. [83] developed an application that interfaces audiobook consumption with other modalities of emphasizing emotion as identified by OpenVokaturi in the speech signal.[7] D'Errico and Poggi [84] use OpenVokaturi to quantify voice-emotion features and iMotions to quantify facial emotion expressions in audio–visual recordings as part of an effort to analyse the distribution of emotions expressed by politicians evidently endeavouring to express humility. Schmidt et al. [85] used emotion classifications from Vokaturi to explore relationships between automated emotion recognition systems (they used OpenFace for facial emotion recognition) and measures of system usability in a think-aloud protocol. Salutari et al. [86] use OpenVokaturi for the speech emotion recognition for a robot developed with multi-modal emotion recognition capacities. These works demonstrate that judgements made by OpenVokaturi in assessing speech emotion are widely adopted.

The openSMILE system is typically used to select a range of low-level acoustic features that have been associated with accurate emotion classification to train emotion classifiers on new datasets. For example, [87] build a model of group emotion

[4] See https://vokaturi.com/algorithms/measuring-emotions — last verified May 2023.

[5] See https://vokaturi.com/algorithms/measuring-acoustic-features — last verified May 2023.

[6] See https://vokaturi.com/algorithms/measuring-emotions — last verified May 2023.

[7] Their user testing participants (n = 6) "criticize that the emotion recognition seems arbitrary at times" [83, p. 866].

from a basis of self-reports of emotion obtained with groups of Mandarin speakers. OpenEar involves default feature set selections and pre-trained emotion classification models derived from openSMILE feature individuation, and has also been adopted by researchers. Golondrino et al. [88] extract emotion qualities from political advertizement jingles, and [89] use a similar approach for assessing educational content. Smith et al. [90] describe the use of OpenEar emotion classification as a component of an approach to automatic composition of movie trailers. The default low-level voice features assessed by openSMILE and the pre-trained models for classifying speech emotion built upon those features are thus also widely adopted.

2.4.2. Facial emotion

A number of researchers have used the emotion classifications produced by facial emotion recognition systems as reliable indicators of emotions experienced by humans pursuing activities in a variety of contexts. Here we note some of these topics.

Novielli et al. [91] use the facial emotion classifications produced by Affectiva as a gold standard in evaluating the classification of wearable biometric sensors. Liu et al. [92] rely on Affectiva to inform analysis of the distribution of emotions experienced by automobile drivers. Sarsenbayeva et al. [93] use the emotion classifications of Affectiva in the analysis of causal relations between smartphone use and emotional experience. Zhou et al. [94] rely on emotion classifications produced by Affectiva in analysing work group structure preferences in design tasks. Hammann et al. [95] use Affectiva's facial emotion classification in research into methods of assessing emotional competence among people with intellectual disabilities. Garcia-Garcia et al. [96] and Singh and Dewan [97] independently use Affectiva's classifications in applications designed to help autistic children recognize and convey facial emotions. As noted above, [69] compared Affectiva iMotions with facial electromyography (EMG) data. They focused on happy, angry and neutral emotions, for each looking at differences between contrast states (e.g. joy-angry/joy+angry) as measured within by the two approaches, normalized to the same scale. A significant correlation is reported between Affectiva and EMG measurements of joy and brow-furrows.

Park and Ryu [98] use the facial emotion classifications of Emotient in an evaluation of teacher acceptance of teacher training systems that simulate scenarios encountered by teachers. Schmitz-Hübsch and Fuchs [99] analyse the relationship between experienced emotions and quality of performance in command and control situations, using Emotient as a facial emotion classifier: finding little in the way of expected interactions (e.g. negative emotions correlating with poorer performance), they reason about the suitability of tasks for eliciting expected ranges of emotions and individual differences in emotional expression. Moreno and Woodruff [100] use Emotient's classifications of facial emotion expression in the evaluation of the effects of background music on the learning experience of students. Davis et al. [101] use the facial emotion classifications of Emotient in order to analyse the role of gender differences in advantages associated with facial expression in bids for micro-lending investments. Trevisan et al. [102] use iMotions Emotient FACET in order to classify emotions in recordings of children with autism and alexithymia, finding that the latter rather than the former impinges on the quantity of emotions expressed. Gupta et al. [103] use iMotions Emotient FACET to classify emotional expressiveness among individuals at high risk of psychosis, and use this in a quantification among blunting of emotions in that group (separately, [104] evaluate the efficacy of iMotions Emotient FACET in relation to human emotion categorization and categorization by an alternative FER (FaceReader, [105]), finding strong agreement for joy). Fischer et al. [106] use Emotient as a means of assessing the outcomes of face transplants, with focus on the tracking of action units rather than emotion classification.

2.4.3. Observations

We have noted applications that depend on speech emotion classification of openSMILE/OpenEar and Vokaturi and facial emotion classifications by Affectiva and Emotient. It is natural to reflect on works that adopt one system or another for each modality in this fashion and whether they would have obtained the same overall conclusions and effects had they adopted the alternative. Our work addresses this by comparing systems as used with defaults and pre-trained models on data sets of recordings made for purposes other than emotion expression, but assembled on the basis of potential emotive content associated with the original purpose, that is, the reason for the recorded speech or press conference. Of course, one could also analyse the interior algorithms and data modelling assumptions of the systems as well as input–output behaviours, and we address these matters (see Section 4.2 and Section 4.3). However, our initial interest here is their classification judgements in relation to data outside their training data sets, the extent to which those judgements are the same or different with respect to shared stimuli and factors that appear to influence differences — cross calibration of the systems.

2.5. Prior comparisons of non-verbal emotion classification systems

Others have also explored comparisons among these available systems, as described below.

2.5.1. Speech emotion

Garcia-Garcia et al. [24] provide a review of available emotion recognition systems, including Vokaturi (and Affectiva, among facial emotion recognition systems), presenting details visible from the literature and system provider websites, but without direct empirical comparison except of the quality of ease of use. Anjum [107] appears to adopt both openSMILE emotion classification and Vokaturi classification, relying most on the latter, but without detailed rationale. Özseven and Düğenci [108] compared openSMILE and Praat [71] as baseline systems using standard datasets, focusing on measurements of acoustic features and overall emotion classification accuracy. Equal accuracy between two systems on datasets with gold-standard labelling does not entail agreement on instances datasets disjoint from training data. This, and the work of Datta et al. [32], described in more detail below (see Section 3.3) are the only efforts of which we are aware that pursues direct comparison of emotion classification with openSMILE and OpenVokaturi pre-trained models.

2.5.2. Facial emotion

Dupré et al. [109] compare Affectiva/Affdex with Emotient /Facet as well as CrowdEmotion's FaceVideo, Microsoft Cognitive Services, MorphCast's EmotionalTracking, EmotionRecognition of Neurodata Lab, VicarVison's FaceReader and VisageTechnologies' FaceAnalysis, addressing databases of posed and spontaneous emotion expression, with reference to independent human judgements. In analysing accuracy in emotion labelling using area under the curve (AUC), both Emotient and Affectiva had a score of 0.77 (no significant difference) for spontaneous emotions and for posed emotions, Emotient had an AUC score of 0.75, while Affectiva's AUC score was 0.79 (difference not significant). Yang et al. [110] evaluate Affectiva in relation to four other systems (Amazon Rekognition, Baidu Research, Face++ and Microsoft Azure) in emotion labelling on standard datasets and then using image distortions (rotation, occlusion, blur, brightness); Affectiva did not emerge as "best" in any of the conditions analysed. Bernin et al. [111] compare Emotient, Affectiva and two other systems (InSight and CERT) with respect to established emotion-labelled

video datasets.[8] Depending on the dataset and emotion, Emotient and Affectiva each at times provide the most correct labels for the relevant category of videos (at times, it is the other two). Correlations across the data sets are not computed (although system classification of an individual item is illustrated), nor are differences subjected to significance testing.

Where possible, we also indicate the extent of cross-system agreement on emotion labelling where they agree on certain underlying physical measurements.

3. Our comparisons of extant non-verbal emotion recognition systems

We briefly outline a methodology for comparing the outputs of two FERs or two SERs given the same input. We then outline the criteria for selecting input data. We will then describe outcome of comparing two SERs with each other, and then two FERs with each other which will include not only the facial emotions but head movements as well. This is followed by a discussion of the methods, input data, and the results of comparison. Initial results of the comparison for facial emotions [31], for speech emotion [32], and head movement [33] are presented. Additional analyses of system agreements and disagreements are also provided.

3.1. A method of comparing two emotion recognition systems

The method we offer for comparing two emotion recognition systems, whether as expressed by voice or face, is to quantify agreement and disagreement on the same data. Essentially, we apply repeated measures to the same samples, and inspect differences in those repeated measures. The systems work with an inventory of emotion labels, and we are therefore inspecting agreement and disagreement in the most likely labels for a given sample. Here we do not compare them on "gold-standard" data, because, as we argue, there is reason to have doubt about "gold-standard" labels themselves.[9] Quantifying agreement and disagreement between systems in repeated measures of the same data provides positive information. Thus, we treat the systems we compare as one would treat any assessment of multiple annotations on a data set. Cohen's κ provides an index of agreement beyond chance, and in the first instance we use this as a measure of system agreement on emotion labels. We then seek more details about the loci of agreement and disagreement by testing the contingency tables of cross-classifications inherent in confusion matrices with a χ^2 test. Significance in such a test means that there is a non-random interaction between the labelling of one system when compared with the labelling of the other system for the same samples. However, a non-random interaction of the labellings does not entail agreement — for example, where one system locates "disgust", the other might systematically find "anger" most likely. Therefore, we inspect the Pearson residuals (a standardized measure of the difference between the observed values and the values that would be expected in each cell of the contingency table if there were no interaction between the column labellings and row labellings).[10] The extent of non-random agreement between the systems is captured in the residuals along the same-label diagonal having greatest magnitude, with positive values, and at a magnitude that signifies significance, but without other cells in the same row or column having greater positive values. This supports identification of significance in "confusions" between the two systems, as well as significant agreement.

For each emotion label that the systems have in common within a modality, we also compute rank-order correlations between their estimates of evidence/confidence in the label for each sample. We use rank-order correlations to avoid distortions that can arise for Pearson correlations when data is not normally distributed. This analysis does not depend on the label with most support for any sample.

In the case of face emotion systems, we also inspect system measurements made that are at a lower-level of abstraction than emotion classification. In particular, we focus on system estimates of yaw, pitch and roll in head movements. It is informative to take these estimates as vectors and we use those vectors to make comparisons. For example, we use cosine between the vectors supplied by each of the systems for a common sample and we also compute for each system, the cosine between the vector between a frame and its preceding frame and compare the resulting cosine values between the two systems. We also consider aggregations such as standard vector magnitudes and pose-angle-sums (the sum of the absolute values of each of yaw, pitch and roll). With these values, we compute correlation coefficients in order to establish levels of agreement, and Wilcox tests to determine whether differences between the systems are statistically significant.

Eventually (see Section 4), we show that it is possible to use system agreement and disagreement constructively as a variable in order to inspect factors that interact with system agreement and disagreement.

A summary indication of the tests we apply in order to quantify judgements is provided in Table 1. Note that the use of the X^2 test for examining interactions of categories we are exploring "confusion matrices" as supplied by compared systems: a significant result does not directly quantify similarity of category assignments by the two systems or difference (although, if the test result is not significant this implies that the null hypothesis must be accepted, in this case, that there is no interaction between the emotion labels provided by one system and the emotion labels provided by the other system, and such a lack of systematicity is tantamount to disagreement). When we inspect the residuals, as indicated above, we can note classification agreement with significant positive residuals along the diagonal and confusion with significant positive residuals elsewhere.

3.2. Acquiring input data for testing emotion recognition systems

In the following we give the criteria for finding videos with their soundtracks:

3.2.1. Source of videos

We have chosen public-domain videoed speeches, with soundtrack, of people who are used to public speaking We select videos shot in a TV studio, or videos shot by professionals for wider distribution,say, in an election campaign.

3.2.2. Typology of our videos

Our videos comprise people who are neither professional actors (or subjects performing according to a prescribed script), nor are these people belong to the general public whose videos can be regarded as videos in the *wild*. We have selected people who have perhaps rehearsed what they are going to say on camera and have learnt to control or exaggerate their emotions in public and

[8] Evidently, CERT [112] provided a foundation for Emotient [28].

[9] Firstly, it is impossible, even for an authentic emoting agent to reliably reconstruct exact emotions from reviewing and labelling each frame or audio segment of their own expressions. Secondly, acted emotions are, by definition, not genuine.

[10] Absolute values between 2 and 4 are significant ($p < 0.05$) and greater than 4 are highly significant ($p < 0.001$); positive values indicate that the count for the cell of interaction exceed what would be expected by chance, and negative residuals indicate that the count for the cell of interaction are less than would be expected by chance.

Table 1
Summary of non-parametric statistical tests that we apply to quantify system similarities and divergences.

Test	General purpose	Compared data types
Wilcox	Differences	Numeric values with respect to two categories
Kruskal	Differences	Numeric values with respect to a more than two categories
Spearman correlation	Similarity	Numeric value agreement
Cohen κ	Similarity	Cross-categorization agreement
χ^2	Interaction	Cross-categorization interaction
χ^2 Pearson residuals	Interaction	Cross-category effect localization

Table 2
Facial emotion analysis — Our video data-base. Counts of individuals by profession and gender; Age ranges; Counts of videos; counts of frames analysed. (POL = Politician; SPO = Spokespersons; CEO = Chief Executive Officer).

Race	Profession			Age		Gender		#Videos		Frames analysed	
	(POL	SPO	CEO)	(Min	Max)	(F	M)	(F	M)	(F	M)
East Asian	9	3	0	49	73	3	9	14	21	46413	155583
South Asian	6	1	3	28	72	1	9	2	24	8094	149056
White European	16	6	7	41	91	14	15	56	45	310594	277572
Total All Races	31	10	10	49	73	18	33	72	90	65101	582211
Total		51				51		162		947312	
Total Time (milliseconds)										31261296	
Total Time (h)										8.68	

Table 3
Nationalities represented among subjects of videos analysed, with counts of Female (F) and Male (M) subjects (F | M).

Race	Nationality (F | M)
East Asian	China (3 | 6), Japan (0 | 1), South Korea (0 | 2)
South Asian	India (1 | 7), Pakistan (0 | 2)
White European	France (1 | 0), Germany (1 | 1) , Ireland (0 | 1), Italy (0 | 1), New Zealand (1 | 0), United Kingdom (1 | 1), United States (10 | 11)

whose speeches are heard and seen by the public at large together with their facial and voice emotions, and head movements. Our database contains *semi-trained actors* and comprise professional politicians, government spokespersons, and CEOs.

3.2.3. Attributes of a selected person

We have chosen three racial groups - *East Asians*, *South Asians*, and *White Europeans*. We have chosen people who are used to speaking in the public, for instance, politicians, their spokespersons, and CEOs (only for FERs). We have tried to have a gender balanced video corpus but were not quite successful due to the presence of the *glass ceiling* against women.

3.2.4. Numbers of video per person

We have endeavoured to have 3 or more videos for each of person in our video base.

3.2.5. Duration of the video

We cannot control the length of public speech but on average our video databases are just over 3 min in duration — this yields a substantial number of frames for analysis as the cycle rate of the FERs ranges between 33–40 frames per milliseconds.

3.2.6. Our video and speech databases

We have used two video databases: for facial emotion recognition we have collected 162 videos of 51 different public speakers from three racial groups in 12 countries (18 females and 33 males) with ages ranging from 49 years to 73 years. The total duration of the 162 videos is 8.68 h (See Tables 2 and 3).

Our second video database comprises 258 videos of 71 different public speakers from three racial groups in 11 countries (25 females and 46 males) with ages ranging from 32 years to 85. The total duration of the 258 videos is 16.66 h. We only use the sound tracks of this collection to compare speech emotion recognition systems; audio clips are sampled at 2000 ms (See Tables 4 and 5).

3.3. Comparing two voice emotion recognition systems

Analysing the application of openSMILE and OpenVokaturi to voice data collected "in the wild" has revealed that those systems reach significantly different conclusions about likely emotions over matched samples. We have used 258 video recordings and associated sound track of the politician talking (16.66 h long) (video plus sound track depicting 64 politicians and 7 official spokespersons. Recordings were processed using a sampling rate of 2000 ms. The data set is described in (Tables 4, 5). We have used three of the US Bureau of Census'racial categories: White Europeans, South Asians, and East Asians. We are collecting data from other categories like African Americans and Africans.

Tests of whether evidence for each emotion (joy, anger, fear, sadness, neutral) was different between the two systems across recordings were conducted. There was a significant difference between the two systems for each of the emotions. Another method of assessing similarity of the systems is determining whether the measurements they produce are sensitive to independent conditions on the data produced. The independent conditions correspond to binary categorizations appropriate to the politicians: Asian (n = 46) vs. White European (n = 25); Female (n = 25) vs. Male (n = 46); under 60 years of age (167 recordings) vs. 60 years of age or older (91 recordings). For nearly every 'emotion and for all three of the binary distinctions, the evidence of the emotion was significantly different between the two categories of the distinction, for both OpenEar and OpenVokaturi. The exceptions were that a significant difference in evidence of neutral depending on the age category was found for OpenEar but not OpenVokaturi, and a significant difference in evidence of anger was identified using the classifications of OpenEar but not OpenVokaturi. Thus, on this level of analysis, much system agreement about relevant distinctions is visible.

Next, we present the analysis of the 258 recordings noted above and agreement on emotion labelling therein.

Firstly, we consider the labelling of the most likely emotion supported by each system. For each of the 2000 millisecond

Table 4
Speech emotion analysis — Our speech audio data-base. Counts of individuals by profession and gender; Age ranges; Counts of videos; counts of clips analysed. (POL=Politician; SPO=Spokespersons; CEO=Chief Executive Officer).

Race	Profession			Age		Gender		#Audios		Clips analysed	
	(POL	SPO	CEO)	(Min	Max)	(F	M)	(F	M)	(F	M)
East Asian	16	3	0	42	71	4	15	12	39	1579	6290
South Asian	17	1	0	32	85	5	13	18	34	1703	3970
White European	31	3	0	32	80	16	18	79	76	7758	8692
Total All Races	64	7	0	32	85	25	46	109	149	11040	18952
Total		71		32	85	71		258		29992	
Total Time (milliseconds)										59984000	
Total Time (hours)										16.66	

Table 5
Nationalities represented among subjects of speech audio recording analysis, with counts of Female (F) and Male (M) subjects (F | M).

Race	Nationality (F | M)
East Asian	China (4 | 11), Japan (0 | 2), South Korea (0 | 2)
South Asian	Bangladesh (1 | 1), India (4 | 10), Pakistan (0 | 2)
White European	Germany (1 | 1) , Ireland (3 | 5), New Zealand (1 | 0), United Kingdom (2 | 2), United States (9| 10)

Table 6
Cross classification of samples using the most likely emotion label according to Vokaturi (columns) in relation to the most likely label according to openSMILE (rows): the confusion matrix diagonal values are in **bold**. The values are counts of clips in each cell of the cross-classification.

		Vokaturi				
		Anger	Fear	Happiness	Neutral	Sadness
openSMILE	anger	**418**	44	304	40	26
	fear	72	**39**	140	22	13
	happiness	56	54	**196**	20	21
	neutral	4266	90	2551	**5058**	1128
	sadness	4077	156	2807	5656	**1888**

samples, both systems provide estimates of confidence in each of the emotion labels. It is natural take the label with maximum confidence for a sample as the system's classification of that sample. In 850 cases, Vokaturi produces "ties". Ignoring ties, the remaining 29,142 samples can be inspected to assess the agreement between the labels with maximum confidence according to each of the systems. To begin, we consider the labels thus derived, and assess agreement between the two systems using a standard measure of inter-annotator agreement: Cohen's weighted κ is estimated as 0.063 — very slight agreement. Examining the cross classification of samples between the most likely labels according to each of the systems offers insight into the locus of agreement and disagreement. Table 6 provides a confusion matrix that results from this cross-classification. To quantify the differences revealed by the confusion matrix, we view the matrix as a contingency table and conduct a Chi-squared test to test whether there is a significant interaction between the row labels and column labels. As the interaction is significant ($\chi^2 = 2244.5, df = 16, p < 2.2e-16$), we inspect Pearson residuals to identify the locus of interacting effects — the corresponding residuals are provided in Table 7. The diagonal values are each significant ($p < 0.05$) and are highly significant ($p < 0.001$) for all but the "neutral" row. However, the diagonal does not contain the greatest positive Pearson residual for each emotion label. Vokaturi estimates of "fear" that coincide with openSMILE estimates of "happiness" are observed even more often than would be expected with no interaction. The divergence between observations and expectations that would follow random interactions for Vokaturi estimates of "neutral" and openSMILE estimates of "sadness" is greater than the divergence between observation and expectation for the diagonal cell, where the systems agree.

For each of the emotion labels that both openSMILE and Vokaturi use, we calculate rank-order correlations in the systems' estimates of applicability of those labels to samples. Thus, we obtain Spearman's ρ values as follows: anger, 0.231 ($p < 0.0001$); fear, 0.128 ($p < 0.0001$); happiness, 0.297 ($p < 0.0001$); neutral, 0.114 ($p < 0.0001$); sadness, 0.081 ($p < 0.0001$). Thus, it can be seen that for no label is there a tremendous level of agreement between the systems about the support for the label, even without reference to system (dis)agreement on the label with maximum support.

3.4. Face and head movement

3.4.1. Facial emotion classification

For each frame, the evidence recorded by Emotient and Affectiva in relation to the emotions joy, fear, disgust, contempt, sadness, surprise, and anger was compared using a rank order test. For the emotions surprise and anger, the null hypothesis that the systems recognize the same level of emotion evidence could not be rejected. For the other emotions, significant differences were noted.

Using the three binary classifications described above, the question was asked whether the emotion evidence was significantly different within the classification, according to each of the three systems. For example, there was a significant difference in the evidence of joy depending on whether the video subject was a female politician or male politician, for all three systems. For all three systems, the gender categorizations also produced significant differences in evidence of anger. For the other emotion labels, significant difference in the emotion evidence according to the gender category appeared in relation to the evidence provided by some systems, but not others. The contrast was visible for both Emotient and Affectiva for five of the seven emotions (and was visible for Azure, only for joy and anger). In relation to the distinction between White Europeans and East Asians, significant

Table 7
Pearson residuals from χ^2 test of the cross classification of samples (recorded in the contingency table in Table 6) using the most likely emotion label according to Vokaturi (columns) in relation to the most likely label according to openSMILE (rows): the confusion matrix diagonal values are in **bold**.

		Vokaturi				
		Anger	Fear	Happiness	Neutral	Sadness
openSMILE	anger	**10.308570**	9.999366	10.145076	−15.277935	−6.596741
	fear	−1.631335	**18.177258**	10.575102	−8.155992	−3.128286
	happiness	−4.844785	23.151010	**14.741508**	−9.574022	−2.582058
	neutral	4.309245	−6.256815	−2.770070	**2.979998**	−6.832353
	sadness	−5.569461	−2.576532	−3.553283	3.444500	**8.885671**

difference in emotion evidence was evident for the estimates of all three systems for contempt, sadness, and anger. In the case of disgust evidence, measurements produced by Emotient and Affectiva were significantly different according to the distinction. Emotient's measurements for evidence of each emotion were significantly different in relation to the distinction. For the binarized age category, there was little agreement in system measurements — evidence of fear as noted by Emotient and Azure were significantly different between the two age categories; for none of the three systems did measurements of disgust or sadness differ significantly between the two age categories.

Finally, direct pairwise system (Spearman) correlations were tested for evidence of each emotion. For Emotient and Affectiva correlations, anger evidence was moderately correlated ($0.3 \leq \rho < 0.6$) and fear evidence demonstrated zero correlation ($0 \leq \rho < 0.1$); evidence of all other emotions were weak ($0.1 \leq \rho < 0.3$). In the Emotient-Azure correlations of emotion evidence, joy, anger, surprise and sadness were moderately correlated and the rest were weakly correlated. In the Affectiva-Azure correlations, anger evidence showed moderate correlation and surprise showed zero correlation, and the rest were weakly correlated.

This work suggested that speech systems required similar analysis (as described above in Section 3.3). The fact that differences in emotion classifications were so prevalent, it also suggested examining whether systems agree on measurements of physical behaviours underlying emotion classification (as described below in Section 3.4.2).

3.4.2. Head movements

Facial emotion recognition systems depend on the analysis of facial landmarks and patterns of combined movement among those landmarks (facial action units). While we intend to examine correlations on these measurements directly, given that we here take a step back from emotion classification by available systems into their measurements of physical quantities that inform emotion classification, we note here that the visibility of landmarks is directly impacted by the movements of the head [33]. We examined the measurements of yaw, pitch and roll estimated by Affectiva and Emotient. We have used the following dataset for the head movement recognition: We have 162 videos of 51 people (18 female; 33 male) which were analysed. We have 31 politicians, 10 CEOs of multinational companies and 10 were professional spokespersons. The videos were observed found on YouTube by searching for prominent individuals providing press conferences and speeches. They were pre-processed to select segments including only the video subject and trimmed to have the face of the subject occupy as much of the frame as possible. Only frames that were registered by both systems were analysed.

Basic measurements of head movement include yaw, pitch and role, and agreement between Affectiva and Emotient on estimates of these angles shows Pearson correlations of 0.857 ($p < 0.001$) for yaw, 0.694 ($p < 0.001$) for pitch and 0.816 ($p < 0.001$) for roll. These results were calculated on a dataset of speeches by politicians, corporate chief executives, and professional press spokespersons from East Asia, South Asia, Europe, the US and Oceana. As we synthesize the work of the earlier paper here, we analyse that data in additional ways, noting, for example, the quartiles in the difference for the two systems on each of those angles (see Table 8). For each video frame, a measurement in degrees of each of yaw, pitch and roll is recorded, and Table 8 shows descriptive statistics of the tendencies in the difference between the Affective measurement (e.g. "A.Yaw") and the Emotient measurement (e.g. "E.Yaw") in each case. This provides a summary description of the scale of difference between the two systems in measuring these low-level physical properties.

The values of yaw, pitch and roll for each frame where both systems recorded measurements were analysed by using two aggregation methods. The first considered "pose angle sums" (PAS) — the sum of the absolute value of each of yaw, pitch and roll. This is similar to vector magnitudes: the square root of the sum of the squares of each vector component. The second considered the value of 1 minus the cosine of the vector of yaw-pitch-roll measurements at a frame and its preceding frame (YPR.df). The measurements provided by Affectiva and Emotient aggregated as PAS values showed a significant positive Spearman correlation ($\rho = 0.62$), but with greater values arising from Affectiva's values than Emotient's (tested with a directional Wilcox test). The Spearman correlation between Affectiva's YPR.df values and those of Emotient (which depends on the cosine of the yaw-pitch-roll vector at a frame and its preceding frame) was moderate ($\rho = 0.33, p < 0.001$). As an additional comparison in the present paper, not reported by Kidambi Murali et al. [33] we note that the Spearman correlation between the YPR vector magnitudes as computed for Affectiva and Emotient is comparable to that of the PAS values: $\rho = 0.63, p < 0.001$.[11] Thus, on the underlying measurements, positive correlations are identifiable but vast differences as well.

The relationship between head movement and emotion classifications of the two systems were explored by Kidambi Murali et al. [33] using these same quantities. For both the PAS and YPR.df values, the means for frames classified by the distinct emotions (anger, contempt, disgust, fear, joy, sadness, surprise) varied significantly and for both systems. For both PAS and YPR.df values, interactions between emotion label and gender were significant for Emotient and Affectiva. Here, we note the distribution of YPR vector magnitudes according to the emotion label reckoned as most probable by each system (see Table 9). The comparison reaches beyond agreement on low-level physical quantities into the more abstract labels inherent in emotion attributions. This suggests the additional analyses that we report in the next section (Section 3.4.3)

3.4.3. Facial emotion in a larger dataset

Above (Section 3.4.1), we addressed comparisons of facial emotion recognition as reported by Ahmad et al. [31]; here we present a comparison as arises in the dataset of Kidambi Murali

[11] Spearman correlations between PAS and YPR magnitudes for Affectiva ($\rho = 0.98, p < 0.001$) and Emotient ($\rho = 0.97, p < 0.001$) are very strong.

Table 8
Quartiles for differences (Δ) between Affectiva and Emotient measurements of Euler angles.

Δ	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
A.Yaw-E.Yaw	−60.780	−5.183	−1.241	−1.150	3.173	52.067
A.Pitch-E.Pitch	−71.3634	0.2504	3.9431	4.4996	8.2334	48.2324
A.Roll-E.Roll	−56.7320	−3.3565	−0.8123	−0.7709	1.7237	85.0020

Table 9
Mean of Yaw-Pitch-Roll (YPR) vector magnitudes for each system and emotion.

System	Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise
Affectiva	23.862	19.494	15.574	12.212	14.977	14.464	14.692
Emotient	11.221	12.309	11.304	10.286	10.902	11.469	11.410

Table 10
Cross classification of frames using the most likely emotion label according to Affectiva (columns) in relation to the most likely label according to Emotient (rows): the confusion matrix diagonal values are in **bold**. The values are the counts of frames that fall into each cell of the cross-tabulation.

		Affectiva						
		Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise
Emotient	anger	**39423**	2215	29030	333	193	9355	7824
	contempt	23110	**4228**	55633	1256	2126	1654	38090
	disgust	51650	2397	**125605**	1366	3409	17994	48438
	fear	19178	1025	36448	**4247**	1636	4024	56010
	joy	25222	6074	56108	1945	**30555**	1529	73547
	sadness	24863	1780	24737	697	798	**7759**	26236
	surprise	14831	831	22483	2887	790	1119	**45552**

Table 11
Pearson residuals of cross classification of frames using the most likely emotion label according to Affectiva (columns) in relation to the most likely label according to Emotient (rows): the cross-classification diagonal residuals are in **bold**. The residuals are derived from the values observed in Table 10.

		Affectiva						
		anger	contempt	disgust	fear	joy	sadness	surprise
Emotient	anger	**156.310**	12.191	−18.103	−24.547	−57.164	84.520	−117.759
	contempt	−18.459	**36.168**	44.589	−10.244	−42.617	−53.724	−4.164
	disgust	−1.129	−35.290	**112.205**	−34.070	−68.178	62.113	−104.134
	fear	−38.828	−27.668	−39.348	**64.891**	−48.073	−20.549	93.519
	joy	−75.290	37.429	−56.647	−12.682	**251.131**	−77.745	54.545
	sadness	51.376	2.398	−39.276	−13.456	−46.512	**60.900**	−3.485
	surprise	−25.716	−21.312	−54.748	49.909	−47.323	−45.665	**110.406**

et al. [33], but not analysed by them. The means of YPR vector magnitudes according to each emotion for the two systems as reported in Table 9 must be contextualized with emphasis on the fact that the frames deemed most likely to be characterized by "joy" for Affectiva need not be the same frames for which "joy" was the most likely label according to Emotient. Firstly, we note that Cohen's unweighted κ is estimated at 0.13 ("slight" agreement). A "confusion" matrix of the classifications deemed most likely by each system is provided in Table 10. It is helpful to quantify the differences inherent in a confusion matrix through inspection of Pearson residuals of a Chi-squared contingency table test ($\chi^2 = 228007, df = 36, p < 2.2e-16$) – the corresponding residuals are provided in Table 11. The significance of the χ^2 statistic allows rejection of the null hypothesis that the there is no interaction between classifications by Affectiva and Emotient, and residuals along the diagonal indicate that the instances of agreement are far greater than would be expected if the relationship between Affectiva and Emotient were random.[12] However, the diagonal does not contain the residuals of greatest positive magnitude in each row and column for contempt, fear or sadness. An alternative comparison of the two systems on each frame considers, for each emotion, the level of evidence deemed available. Examining the Spearman correlations between Emotient and Affectiva for the estimates of intensity of the basic emotions, one finds for anger, $\rho = 0.310, p < 0.001$; for joy, $\rho = 0.228, p < 0.001$, for sadness, $\rho = 0.071, p < 0.001$, for disgust, $\rho = 0.061, p < 0.001$, for fear, $\rho = -0.002, p = 0.06625$ (no significant correlation), for surprise, $\rho = 0.216, p < 0.001$. If one restricts the analysis to those frames for which the difference between Affectiva's and Emotient's estimates of each Euler angle was a value between its first and third quartile (see Table 8),[13] then, for anger, $\rho = 0.542, p < 0.001$; for joy, $\rho = 0.373, p < 0.001$, for sadness, $\rho = 0.235, p < 0.001$, for disgust, $\rho = 0.230, p < 0.001$, for fear, $\rho = 0.138, p < 0.001$, for surprise, $\rho = 0.419, p < 0.001$. That is, in each case, the correlation between the systems' judgements of emotions is stronger for the frames for which the systems agree on the measurements of the underlying head movements, even if those correlations do not reach overwhelming agreement.

[12] The absolute value of the Pearson residual indicates significance: absolute values between 2 and 4 are significant ($\alpha = 0.05$) and absolute values greater than 4 are highly significant ($\alpha < 0.001$). The sign indicates the direction of divergence from random interaction expectations: positive values indicate that observations exceed expectations under no interactions; negative values indicate fewer observations in a cell than would be expected with no interactions.

[13] Within this restriction on the dataset, the Pearson correlations between Affectiva and Emotient estimates of yaw, pitch and roll become 0.967 ($p < 0.001$), 0.936 ($p < 0.001$) and 0.957 ($p < 0.001$), respectively.

Table 12
Frame counts by sex of person depicted and agreement between Affectiva and Emotion on the most likely emotion.

System agreement	Sex of person depicted in the frame	
	Female	Male
Most likely emotion labels disagree	282729	418142
Most likely emotion labels agree	82372	174997

3.5. Summary

This section has used correlation analysis to identify the extent of agreement between voice and facial emotion recognition systems on classifications they provide with numeric values for high-level judgements (like confidence in or evidence for emotion labels) and judgements of lower-level quantities (like head movement). We also use categorial cross-classification tests of agreement. We note some level of agreement, but, more importantly, also disagreement. We demonstrate that the disagreements are statistically significant, for both high-level and low-level classifications. In the next section, we show that system disagreement may be used effectively as a response variable in order to identify factors that interact with disagreement, and we interpret the results reported in this section more generally.

4. Discussion

4.1. General observations

We have demonstrated that researchers have made use of voice and facial emotion classification systems using their default models in order to inform other processes. This creates the question of whether such systems are interchangeable. We have shown that they are not: prominent available speech and facial emotion recognition systems differ in their classifications of emotions within their modality of application.[14] Further, we have shown that available facial emotion recognition systems differ in assessments of physical quantities that underpin emotion recognition. The work we synthesize here has demonstrated that in some cases, across systems, similar distinctions are visible within the measurements, for example, interactions of age, gender and ethnicity. However, the extent to which those distinctions truly are congruent is not yet clear.

The implication of these observations is that it is dangerous to use a single emotion classification system in work that depends on emotion classification. It may be necessary to restrict inference to those cases where measurements of constructs at conceptual levels lower than emotion classification can be shown to be agreed by more than one system, when using these systems on data outside the scope of their training sets. Simply using multiple systems and majority classification seems less prudent than identifying the physical conditions in which systems agree and disagree and determining how to relate system disagreements. Our ongoing research in this area at present includes exactly that.

Consider two examples of the analysis of disagreement. Firstly, consider the binary condition of whether the two systems agree on an emotion label in relation to the sex of the person depicted in each frame. This yields a two-by-two contingency table, as in Table 12.

A Chi-squared test of the interactions in the cross categorization reveals significance ($\chi^2 = 5543.8, df = 1, p < 2.2e-16$). Inspection of the residuals reveals highly significant effects — system agreement on emotion is far less than would be expected for women than if there were no interaction between emotion label agreement and sex and far greater than would be expected for men than if there were no interaction. For system disagreement, the surprise observations are in the other direction: far more observations of disagreement on emotion labels for females than would be expected with no interaction, and far fewer disagreements on emotion labels for males than would be expected with no interaction.

The approach of this example may also be applied to voice emotion recognition.

Table 13 depicts the cross-classification of system agreement and sex. A χ^2 test of the contingency table supports the rejection of the null hypothesis that there is no interaction between sex of the recorded person and system agreement or disagreement ($\chi^2 = 361.45, df = 1, p < 0.0001$). The interpretation of residuals is the same as for system agreement on face emotion labels and sex: there is far less system agreement for females than would be expected of a random interaction and far more system agreement for males than would be expected with only random interaction, and the converse relations apply to system disagreement — significantly more for females and significantly less for males than would be expected without an interaction.

As a second example, consider the measurement of roll, which is at a lower-level of abstraction than emotion labelling, and recall Table 8. One may ask the question of whether the difference between roll as measured by Affectiva and by Emotient is significant as measured on frames depicting men (mean $= -0.7659796$) and frames depicting women (mean $= -0.7789769$). A Wilcox test reveals that the difference is significant ($W = 1.0986e+11, p < 2.2e-16$). Thus, as argued in other contexts [113], examining qualities of the data and their relationships to disagreements in judgements can be illuminating: disagreement in judgement provides a quantity that is as interesting to study as variations in the quantity that underlies the disagreement. Disagreement is itself a useful quantity to study as a response variable in relation to other factors, including the factors that one would want to analyse in relation to the values underlying the disagreement. That is, someone who wants to study the interaction of emotion with sex, age and profession would probably gain by using more than one system to judge emotion, and to also study the interaction of system disagreement on emotion with sex, age and profession. We have shown how this might work for both category labels (probable emotion) and numeric values (angles of roll).

Notice that the "problem" of emotion label disagreement is not solved by embarking on feature extraction and model training for emotion labelling oneself. Fundamentally, the problem is that the primary labelled training data involves simulated emotions or speculation regarding what the data subject was feeling at any moment, and neither sort of label can be deemed to achieve a "gold standard" for the purpose of comparing data gathered from other settings. Thus, there is an inherent epistemic gap in affective computing.

Researchers in other disciplines, those who study human resource management, for example, worry about the ethical ramifications of the functionality of systems that include automatic emotion recognition (see, for instance, work by Fernández-Martínez and Fernández [114]) — and their concerns about fully reliable systems should be amplified given the system-dependent nature of categorizations made.

[14] Recalling the metaphor of twisted fibres from the introduction, we expect disagreement in expressed emotion between modalities at many moments of sampling.

Table 13
Sample counts by sex of person recorded and agreement between Vokaturi and openSMILE on the most likely voice emotion.

System agreement	Sex of person depicted in the sample	
	Female	Male
Most likely emotion labels disagree	8595	12948
Most likely emotion labels agree	2102	5497

4.2. Notes on the differences in the outputs of facial emotion recognition systems

Facial emotion recognition systems (FERS) rely on the accurate identification of movement in certain areas of the face. Especially where two or more parts of the face meet – lips and nose, 'gaps' in the face specifically the mouth and eyes, hairlines around eyebrows and the forehead: areas of different shapes and textures, controlled by specific group of facial muscles – the so-called facial landmarks. The location and movement of these landmarks are correlated with one or more candidate emotions. This ontology of facial emotion expressions relies on muscle movement (cf. action units), the contraction and relaxation of the muscles, and a prescription, based partly on observation that certain muscles movements move more or less during a bout of emotional expression. The landmarks move and then return to their original position. The muscles are sub-dermal and are found universally so if we accept that one or more muscle movements is a physical correlate of a basic emotion irrespective of the social and situational context, then we may have systems like Emotient and Affectiva that may be applied universally. However, The movement of skin covering the face is conditioned by the underlying anatomy — and the anatomy does vary with race/ethnicity; the landmarks are tracked via collections of pixel in video recordings, and the reflectivity of these pixels will be conditioned by the colour of the skin as well. But the results produced by these systems are different for certain number of the basic emotions. We suggest the reasons for this below.

There are many reasons why the outputs of the two systems operating on our database of politicians usually appearing in front of the camera, sometimes interspersed with their profile view. These politicians are of different ages, ethnicities, and cultures. For us the explanation lies in the (i) differences in the architectures and training regimens of Emotient and Affectiva (and differences between openSMILE and Open Vokaturi); (ii) the variation in facial anthropometrics in the real world in terms of race, colour and age, and in vocal attributes anthropometrics in general (facial, laryngeal, and voice articulatory systems that vary across the three main axes; (iii) reliance on assumptions in computer algorithms regarding luminance conservation in optical flow and (iv) the use representation schemes used for facial and vocal features sets that may lack space–time invariance.

4.2.1. The architectural and training regimens difference

The emotient pipeline. The pipeline used to detect facial emotions comprises: Detection of face and facial features, the face is partitioned into many (scaled) segments, image filters are applied to identify facial components and for smoothing — this involves Box, Gabor, and spatiotemporal filters, plus local (pixel) orientation filters, this is followed by the application of statistical machine learning algorithms for 'learning' to select facial features, using Ada Boost and related techniques, finally classifiers are used to make decisions about 'the presence of an action unit' and this is then correlated with the presence of an emotion or emotions; for example, Gabor wavelets of the (emotion-labelled) face are used as an input to a support vector machine for identifying the labelled emotion during training. There is an extensive use of principal and independent component analysis in the emotion identification phase. The authors have used backpropagation algorithms in their work as well [20,115–117]. Emotient was trained on a variation of the Cohn-Kanade data bases [118–120].

The affectiva pipeline. In Affectiva the facial emotion expression analysis platform is similar in architecture to Emotient, but there are many and subtle differences. The video is fed into the pipeline first for face detection and coding, image features are then identified using histogram-oriented gradient features and using scale-invariant feature transformation, for feature classification Affectiva uses SVM and the SVM output is post-processed to remove noise and then output candidate emotion category. The training data set used by el Kaliouby and colleagues contains 15 million or so frames, and they use both 40,000 positive and 40,000 negative examples of each emotion or what they call mental states [21,22,70,121].

There are key differences between the two systems in each of the different parts of the emotion identification pipeline: The way the face is identified, different optical filters are used to identify shapes and for noise removal, the training data sets are different in sizes and in composition — Emotient was trained on a variant of the Cohn-Kanada data base of posed emotions (c. 4000–8000 subjects, [120]) whereas Affectiva is trained on a spontaneously-expressed, very large crowd-sourced [122] video database, albeit the subjects were asked to express an opinion about the stimulus they received.

4.2.2. Variation in facial anthropometrics

There is an implicit assumption in facial emotion recognition literature that basic emotions are universal, so the face deformation required to produce an emotion will be the same irrespective of race, colour and age. There are lessons from related literature in forensic science and restorative surgery which may explain that differences in the facial anthropometrics between our sample and that of the training data used to train Emotient and Affectiva may cause the differences in the results in the two systems. Our sample may contain racial/pigmentation/age outliers not in the training data bases.

This is a big issue in face recognition systems where the face of a 'person of interest' is matched against a data base: essentially key features of a face are to be matched against the pre-stored features. The image of a person of interest is taken in non-ideal circumstances, say, at the scene of a crime where lightning conditions, camera evasion, face covering and other factors tend to produce a low quality image. The users of face recognition systems use both frontal and profile view, and in each case they are looking at, or a computer system, is matching permanent and transient distances between key areas in the face: the permanent horizontal distances include inner/outer eye distances, face/mouth/nose width, and the permanent vertical distances relate to eyeline to nose base/mouth/chin/lips distances and ear height. The transient distances relate to eyeline to hairline/eyebrow distances, right/left eyebrow width and height. Similarly for the profile view. There is a suggestion that a database of the faces of known individual will be 'unreliable unless multiple distance and angular measurements from both profile and full-face images were included in an analysis' [123]. Similarly,

a frontal/view image of human face is taken before restorative facial surgery, and here the consideration always incorporates how the patient will express facial emotions post surgery [124]. We learn from the facial anthropometrics literature that there are 'statistically significant differences [...between] between males and females, [different] racial/ethnic groups, and the subjects who were at least 45 years old when compared to workers between 18 and 29 years of age' [10]. The other related notion in forensic science literature is the effect of facial expressions on facial anthropometrics in that facial expressions sometimes can distort an image to an extent cannot be easily matched to a pre-stored image which either may have no expression or a different expression [125]

4.3. Notes on the differences in the outputs of speech emotion recognition systems

The universality of six or seven basic emotions plus boredom are the ontological basis of many speech recognition systems including openSMILE and Open Vokaturi. There are three reasons for the differences in the results. The ontological basis of the speech emotion recognition systems is the context-independent existence of three physical correlates of speech emotion expression: prosody, spectral distribution of voice attributes, and quality.

4.3.1. The architectural and training regimens difference

The pipelines for both the systems are based on voice identification, feature extraction, and finding physical correlates of the features thus extracted. Both openSMILE and Vokaturi use two prosodic features, pitch and duration, for spectral features both measure formant frequencies. For spectral features openSMILE uses the mel-spectra and for quality Vokaturi uses moments like shimmer and jitter (although openSMILE computes the two quality features and also Harmonic Mean Ratio but it is not used in our version of the software). The training of these systems uses statistical machine learning techniques like principal component analysis and support vector machines. The two systems were on two emotionally labelled speech data bases each and had one common data base: openSMILE was trained on the Berlin Speech Database of young male and female German native-speakers, and Vokaturi was trained on an emotionally labelled data base, SAVEE, comprising the speech of four young English native speakers. Both the systems used e-Interface data base of speakers from 14 nationalities — mainly European citizens. The choice of the features may influence the final results and certainly will impact in the understanding of the output of non-European speakers.

4.3.2. Variation in voice attributes related to anthropometrics

There are well documented cases of 'racial disparities': For instance, take the example of automatic speech recognition and transcription: ASR systems marketed by Amazon, Apple, Google, IBM, and Microsoft — when tested to transcribe the structured interviews of white and black speakers (42:73, totalling around 20 h of transcription). The word error for black speakers was double that of white speakers. It was claimed that the 'disparities [are due] to the underlying acoustic models used by the ASR systems' creating a 'race gap' [51]. In a comparison of vocal tract dimensions and formant frequencies in a gender balanced sample of 40 white Americans, 40 African Americans, and 40 Chinese subjects, researchers recorded first three formant frequencies of nine isolated vowels an measured the vocal tract cavities of all the subjects. Significant race and gender were found both in vocal tract dimensions and in the formant frequencies [126].

4.4. Summary

In this section we have interpreted the ramifications of differences in the input–output behaviours of the SER and FER systems we have analysed. We also highlight differences in the modelling assumptions and algorithms used that may account for some of these differences. The exact relationship between the internal approach of the system and the empirical consequences, apart from cross-system difference, remains open for exploration.

5. Conclusions

In considering the material that we have presented, limitations must be acknowledged. Firstly, it is self-evident that this is work in progress. We have yet to analyse the extent of agreement on measurements of low-level features in speech emotion recognition systems, the tracking of facial landmarks in facial emotion recognition systems, or the full interactions between measurements of physical quantities for facial emotion systems and the other interacting categories. Moreover, the data set that we have been developing through these projects is expanding, firstly to allow more systematic treatment of the categories and their internal structure, and secondly to support additional dimensions of probable contrasts in emotion recognition. However, this implicitly reveals that the samples we have analysed here are relatively small.[15] We have not added independent human assessment of the emotions that might have been experienced by each speaker at each (or any) frame, since our purpose is to calculate the potential for intersubstitution of extant systems applied to data outside their training sets through analysing their agreements and disagreements.[16]

Some immediate next steps are obvious, and we hope that others will join us in the efforts. Firstly, a greater range of the low-level features that are estimated by SERs and FERs may be approached using the same general approach that we have employed here. Secondly, data sets for which independent human annotation of emotion has been applied would be useful to explore using this same method of system comparison, adding the human annotations as another reference point (while acknowledging that it is difficult for human annotation of emotions to achieve a gold standard — acted emotion is not reliably authentic emotion, and *post-hoc* labelling, whether by the emoting subject or third parties, is not reliable, either). Thirdly, re-analysing data sets that may be available where judgements of only one system have been relied upon would enable quantification of the extent to which disagreements would make a difference to follow-on processing in actual systems. Fourthly, larger data sets constructed in the spirit of the one that we have analysed would be useful to explore.

We present this report on the current state of our work in this area (which, although not complete, does provide a coherent encapsulation and synthesis), because we think that others will benefit from seeing the explicit demonstration that the systems that we have addressed are not fully intersubstitutable for each other. Therefore, it is not safe, in general, to simply adopt one of these systems as freely available "off-the-shelf" systems as a component in a larger system, without reflection on the fact that an alternative system might provide different classifications in critical cases. Investigation of the causes of system disagreements and whether they are relevant to the application at hand is important. Knowing their comparative accuracy on standard

15 On the other hand, the data is available to others for independent analysis by contacting the second author.

16 It has been argued elsewhere that analysis of annotation disagreement is itself of theoretical interest [113].

datasets is different to seeing the judgements they make on data "in the wild". Moreover, we emphasize the value of using system disagreement as a response variable as a supplement to using individual system judgements and restricting analysis to areas of agreement. We do not intend for this to be a pessimistic message. Rather, our goal is to identify parameters that determine when they are interchangeable and when they are not. Researchers accept the fact that sometimes parametric tests of inferential statistics are robustly applicable to a dataset, and sometimes only non-parametric tests are safe, depending on properties of the data. Our goal is to have the settings of applicability of emotion recognition systems similarly identified and accepted. This can only enhance research that builds on emotion recognition systems.

Funding

Funding from the Discipline of Artificial Intelligence within the School of Computer Science and Statistics of Trinity College Dublin, the University of Dublin, is gratefully acknowledged.

CRediT authorship contribution statement

Carl Vogel: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Khurshid Ahmad:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors are grateful from the interactions that have arisen from participation in the Gestures and Head Movement (GeHM) research network (Independent Research Fund Denmark grant 9055-00004B). We are grateful to students with whom we have collaborated on this topic: Deepayan Datta, Clodagh Lynch, Dr. Shirui Wang, Wanying Jiang, Yatheendra Pravan Kidambi Murali. We thank Dr. Maria Koutsombogera for helpful feedback on an earlier draft of this paper. The reviewers provided extremely useful constructive criticism of our submission, and this revision is improved by responding to their remarks. Errors, of course, are our own.

References

[1] Rosalind W. Picard, Jonathan Klein, Computers that recognise and respond to user emotion: theoretical and practical implications, Interact. Comput. 14 (2) (2002) 141–169.
[2] Björn Schuller, Anton Batliner, Stefan Steidl, Dino Seppi, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, Speech Commun. 53 (9–10) (2011) 1062–1087.
[3] Deborah M. Riby, Lisa Whittle, Gwyneth Doherty-Sneddon, Physiological reactivity to faces via live and video-mediated communication in typical and atypical development, J. Clin. Exp. Neuropsychol. 34 (4) (2012) 385–395, http://dx.doi.org/10.1080/13803395.2011.645019, PMID: 22260255.
[4] Ray D. Kent, The uniqueness of speech among motor systems, Clin. Linguist. Phonetics (ISSN: 0269-9206) 18 (6–8) (2004) 495–505.
[5] Steve An Xue, Jianping G. Hao, Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry, J. Voice (ISSN: 0892-1997) 20 (3) (2006) 391–400.
[6] Xingyu Chen, Zhengxiong Li, Srirangaraj Setlur, Wenyao Xu, Exploring racial and gender disparities in voice biometrics, Sci. Rep. (ISSN: 2045-2322) 12 (1) (2022) 1–12.
[7] Hyoung-Jin Moon, Won Lee, Ji Yun Choi, Dynamic evaluation of facial muscles: 3D skin displacement vector analysis using a facial painting model, Laryngoscope Investigative Otolaryngol. (ISSN: 2378-8038) 6 (4) (2021) 650–656.
[8] Tim Rawlinson, Abhir Bhalerao, Li Wang, Principles and methods for face recognition and face modelling, in: Handbook of Research on Computational Forensics, Digital Crime, and Investigation: Methods and Solutions, Igi Global, 2010, pp. 53–78.
[9] Kaori Amano, Michiko Naito, Masato Matsuo, Morphological study of human facial fascia and subcutaneous tissue structure by region through SEM observation, Tissue Cell (ISSN: 0040-8166) 67 (2020) 101437.
[10] Ziqing Zhuang, Douglas Landsittel, Stacey Benson, Raymond Roberge, Ronald Shaffer, Facial anthropometric differences among gender, ethnicity, and age groups, Ann. Occup. Hyg. (ISSN: 1475-3162) 54 (4) (2010) 391–402.
[11] Paul Ekman, Wallace V. Friesen, Nonverbal leakage and clues to deception, Psychiatry (ISSN: 0033-2747) 32 (1) (1969) 88–106.
[12] Paul Ekman, Wallace V. Friesen, Phoebe Ellsworth, Emotion in the Human Face: Guidelines for Research and an Integration of Findings, Elsevier, ISBN: 1483147630, 2013.
[13] Marc Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, Stan C.A.M. Gielen, Acoustic correlates of emotion dimensions in view of speech synthesis, in: INTERSPEECH, 2001, pp. 87–90.
[14] Peter Claes, Dirk Vandermeulen, Sven De Greef, Guy Willems, John Gerald Clement, Paul Suetens, Computerized craniofacial reconstruction: Conceptual framework and review, Forens. Sci. Int. (ISSN: 0379-0738) 201 (1–3) (2010) 138–145.
[15] Caroline Wilkinson, Christopher Rynn, Craniofacial Identification, Cambridge University Press, ISBN: 110737684X, 2012.
[16] Thanapoom Boonipat, Jason Lin, Uldis Bite, Detection of baseline emotion in brow lift patients using artificial intelligence, Aesthetic Plast. Surg. (ISSN: 1432-5241) 45 (6) (2021) 2742–2748.
[17] Keon M. Parsa, William Gao, Jack Lally, Stephen P. Davison, Michael J. Reilly, Evaluation of personality perception in men before and after facial cosmetic surgery, JAMA Facial Plastic Surg. 21 (5) (2019) 369–374, http://dx.doi.org/10.1001/jamafacial.2019.0463.
[18] Marie L. Smith, Daniel Grühn, Ann Bevitt, Mark Ellis, Oana Ciripan, Susan Scrimgeour, Michael Papasavva, Louise Ewing, Transmitting and decoding facial expressions of emotion during healthy aging: More similarities than differences, J. Vis. (ISSN: 1534-7362) 18 (9) (2018) 10.
[19] Judith A. Markowitz, Voice biometrics, Commun. ACM (ISSN: 0001-0782) 43 (9) (2000) 66–73.
[20] Marian Stewart Bartlett, Gwen Littlewort-Ford, Javier Movellan, Ian Fasel, Mark Frank, Automated facial action coding system, 2014, US Patent US 8, 798, 374 B2.
[21] Rana El Kaliouby, Rosalind W. Picard, Abdelrahman N. Mahmoud, Youssef Kashef, Miriam Anna Rimm Madsen, Mina Mikhail, Method and system for real-time and offline analysis, inference, tagging of and responding to person (s) experiences, 2011, Google Patents, Google Patents.
[22] Rana El Kaliouby, Peter Robinson, Mind reading machines: Automated inference of cognitive mental states from video, in: Systems, Man and Cybernetics, 2004 IEEE International Conference on, Vol. 1, IEEE, ISBN: 0780385667, 2004, pp. 682–688.
[23] Florian Eyben, Martin Wöllmer, Björn Schuller, OpenEAR—introducing the munich open-source emotion and affect recognition toolkit, in: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, IEEE, ISBN: 142444800X, 2009, pp. 1–6.
[24] Jose Maria Garcia-Garcia, Victor M.R. Penichet, Maria D. Lozano, Emotion detection: A technology review, in: Proceedings of the XVIII International Conference on Human Computer Interaction, in: Interacción '17, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450352291, 2017, pp. 1–8, http://dx.doi.org/10.1145/3123818.3123852.
[25] Huafeng Jin, Shuo Wang, Voice-based determination of physical and emotional characteristics of users, 2018, Google Patents.
[26] Joao Palotti, Gagan Narula, Lekan Raheem, Herbert Bay, Analysis of emotion annotation strength improves generalization in speech emotion recognition models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5828–5836.
[27] Jeffrey F. Cohn, Zara Ambadar, Paul Ekman, Observer-based measurement of facial expression with the facial action coding system, in: The Handbook of Emotion Elicitation and Assessment, Vol. 1 no. 3, 2007, pp. 203–221.

[28] Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, Kang Lee, Automatic decoding of facial movements reveals deceptive pain expressions, Curr. Biol. (ISSN: 0960-9822) 24 (7) (2014) 738–743, http://dx.doi.org/10.1016/j.cub.2014.02.009, URL https://www.sciencedirect.com/science/article/pii/S096098221400147X.

[29] Daniel McDuff, Rana El Kaliouby, Applications of automated facial coding in media measurement, IEEE Trans. Affect. Comput. 8 (2) (2016) 148–160.

[30] Shan Li, Weihong Deng, Deep facial expression recognition: A survey, IEEE Trans. Affect. Comput. 13 (3) (2020) 1195–1215.

[31] Khurshid Ahmad, Shirui Wang, Carl Vogel, Pranav Jain, Oscar O'Neill, Basit Hamid Sufi, Comparing the performance of facial emotion recognition systems on real-life videos: Gender, ethnicity and age, in: Kohei Arai (Ed.), Proceedings of the Future Technologies Conference, Volume 1, FTC2021, in: Lecture Notes in Networks and Systems Volume 358, Cham, Switzerland: Springer International Publishing, 2022, pp. 193–210.

[32] Deepayan Datta, Wanying Jiang, Carl Vogel, Khurshid Ahmad, Speech emotion recognition systems: A cross-language, inter-racial, and cross-gender comparison, in: Advances in Information and Communication: Proceedings of the 2023 Future of Information and Communication Conference, Vol. 1, FICC, Springer, 2023, pp. 375–390.

[33] Yatheendra Pravan Kidambi Murali, Carl Vogel, Khurshid Ahmad, Head orientation of public speakers: Variation with emotion, profession and age, in: Advances in Information and Communication: Proceedings of the 2023 Future of Information and Communication Conference, Vol. 2, FICC, Springer, 2023, pp. 79–95.

[34] Maria Spyropoulou, Khurshid Ahmad, Disaster-related public speeches: The role of emotions, in: 2016 11th International Conference on Availability, Reliability and Security, ARES, IEEE, 2016, pp. 800–804.

[35] Philip J. Stone, Earl B. Hunt, A computer approach to content analysis: studies using the general inquirer system, in: Proceedings of the May 21-23, 1963, Spring Joint Computer Conference, 1963, pp. 241–256.

[36] James. W. Pennebaker, Martha E. Francis, Roger J. Booth, Linguistic Inquiry and Word Count (LIWC), Erlbaum, Mahwah, NJ, 2001.

[37] Khurshid Ahmad, Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology, Vol. 45, Springer Science & Business Media, 2011.

[38] Clayton Hutto, Eric Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 8, no. 1, 2014, pp. 216–225.

[39] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, Antonio Feraco, et al., A Practical Guide to Sentiment Analysis, Springer, 2017.

[40] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, Meeyoung Cha, Comparing and combining sentiment analysis methods, in: Proceedings of the First ACM Conference on Online Social Networks, 2013, pp. 27–38.

[41] Rui Mao, Qian Liu, Kai He, Wei Li, Erik Cambria, The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection, IEEE Trans. Affect. Comput. (2022) 1–11, http://dx.doi.org/10.1109/TAFFC.2022.3204972.

[42] Wei Li, Luyao Zhu, Rui Mao, Erik Cambria, SKIER: A symbolic knowledge integrated model for conversational emotion recognition, in: Proceedings of the 37th AAAI Conference on Aritificial Intelligence, 37, (11) 2023, pp. 13121–13129, http://dx.doi.org/10.1609/aaai.v37i11.26541.

[43] Tahira Reid, James Gibert, Inclusion in human–machine interactions, Science (ISSN: 0036-8075) 375 (6577) (2022) 149–150.

[44] Gizelle Anzures, Paul C. Quinn, Olivier Pascalis, Alan M. Slater, Kang Lee, Development of own-race biases, Vis. Cogn. (ISSN: 1350-6285) 21 (9–10) (2013) 1165–1182.

[45] Kerry Kawakami, Justin P. Friesen, Xia Fang, Perceiving ingroup and outgroup faces within and across nations, Br. J. Psychol. (ISSN: 0007-1269) 113 (3) (2022) 551–574.

[46] Zhilong Guo, Lewis Kennedy, Policing based on automatic facial recognition, Artif. Intell. Law (ISSN: 1572-8382) (2022) 1–47.

[47] J.B. Dawson, D.J. Barker, D.J. Ellis, J.A. Cotterill, E. Grassam, G.W. Fisher, J.W. Feather, A theoretical and experimental study of light absorption and scattering by in vivo skin, Phys. Med. Biol. (ISSN: 0031-9155) 25 (4) (1980) 695–709.

[48] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, Analysis of human faces using a measurement-based skin reflectance model, ACM Trans. Graph. (ToG) (ISSN: 0730-0301) 25 (3) (2006) 1013–1024.

[49] Karan Chopra, Daniel Calva, Michael Sosin, Kashyap Komarraju Tadisina, Abhishake Banda, Carla De La Cruz, Muhammad R. Chaudhry, Teklu Legesse, Cinithia B. Drachenberg, Paul N. Manson, A comprehensive examination of topographic thickness of skin in the human face, Aesthetic Surg. J. (ISSN: 1527-330X) 35 (8) (2015) 1007–1013.

[50] Michael Eggerstedt, Jessica Rhee, Megan Buranosky, Pete S. Batra, Bobby A. Tajudeen, Ryan M. Smith, Peter C. Revenaugh, Nasal skin and soft tissue thickness variation among differing races and ethnicities: An objective radiographic analysis, Facial Plastic Surg. Aesthetic Med. (ISSN: 2689-3614) 22 (3) (2020) 188–194.

[51] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, Sharad Goel, Racial disparities in automated speech recognition, Proc. Natl. Acad. Sci. (ISSN: 0027-8424) 117 (14) (2020) 7684–7689.

[52] Cade Metz, There is a racial divide in speech-recognition systems, researchers say, N.Y. Times (2020).

[53] Joan Palmiter Bajorek, Voice recognition still has significant race and gender biases, Harv. Bus. Rev. Digit. Articles May 10, 2019 (2019) 1–5, URL https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases.

[54] Rachael Tatman, Gender and dialect bias in YouTube's automatic captions, in: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, 2017, pp. 53–59.

[55] Matthew L. Rohlfing, Daniel P. Buckley, Jacquelyn Piraquive, Cara E. Stepp, Lauren F. Tracy, Hey Siri: How effective are common voice recognition systems at recognizing dysphonic voices? Laryngoscope (ISSN: 0023-852X) 131 (7) (2021) 1599–1607.

[56] Michele Merler, Nalini Ratha, Rogerio S. Feris, John R. Smith, Diversity in faces, 2019, arXiv preprint arXiv:1901.10436.

[57] Mehmet Berkehan Akçay, Kaya Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Commun. (ISSN: 0167-6393) 116 (2020) 56–76.

[58] Amjad Rehman Khan, Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges, Information (ISSN: 2078-2489) 13 (6) (2022) 268, http://dx.doi.org/10.3390/info1306.

[59] William James, What is an emotion? Mind 9 (34) (1884) 188–205.

[60] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, Seth D. Pollak, Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements, Psychol. Sci. Public Interest (ISSN: 1529-1006) 20 (1) (2019) 1–68.

[61] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Process. Mag. (ISSN: 1053-5888) 29 (6) (2012) 82–97.

[62] Dacher Keltner, Disa Sauter, Jessica Tracy, Alan Cowen, Emotional expression: Advances in basic emotion theory, J. Nonverbal Behav. (ISSN: 1573-3653) 43 (2) (2019) 133–160.

[63] Rachael E. Jack, Wei Sun, Ioannis Delis, Oliver G.B. Garrod, Philippe G. Schyns, Four not six: Revealing culturally common facial expressions of emotion, J. Exp. Psychol. [Gen.] (ISSN: 1939-2222) 145 (6) (2016) 708–730.

[64] Anna Esposito, Maria Teresa Riviello, Nikolaos Bourbakis, Cultural specific effects on the recognition of basic emotions: A study on Italian subjects, in: Andreas Holzinger, Klaus Miesenberger (Eds.), HCI and Usability for E-Inclusion, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 135–148, http://dx.doi.org/10.1007/978-3-642-10308-7_9.

[65] Daniela Schneevogt, Patrizia Paggio, The effect of gender and age differences on the recognition of emotions from facial expressions, in: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, PEOPLES, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 11–19, URL https://aclanthology.org/W16-4302.

[66] Anna Esposito, Terry Amorese, Nelson Mauro Maldonato, Alessandro Vinciarelli, Maria Ines Torres, Sergio Escalera, Gennaro Cordasco, Seniors' ability to decode differently aged facial emotional expressions, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, 2020, pp. 716–722, http://dx.doi.org/10.1109/FG47880.2020.00077.

[67] Aliki Economides, Yiannis Laouris, Massimiliano Conson, Anna Esposito, Facial emotion recognition skills and measures in children and adolescents with Attention Deficit Hyperactivity Disorder (ADHD), in: Anna Esposito, Marcos Faundez-Zanuy, Francesco Carlo Morabito, Eros Pasero (Eds.), Progresses in Artificial Intelligence and Neural Systems, Springer Singapore, Singapore, ISBN: 978-981-15-5093-5, 2021, pp. 435–475, http://dx.doi.org/10.1007/978-981-15-5093-5_39.

[68] Anna Esposito, Filomena Scibelli, Alessandro Vinciarelli, A pilot study on the decoding of dynamic emotional expressions in major depressive disorder, in: Simone Bassis, Anna Esposito, Francesco Carlo Morabito, Eros Pasero (Eds.), Advances in Neural Networks, Springer International Publishing, Cham, ISBN: 978-3-319-33747-0, 2016, pp. 189–200.

[69] Louisa Kulke, Dennis Feyerabend, Annekathrin Schacht, A comparison of the affectiva iMotions facial expression analysis software with EMG for identifying facial expressions of emotion, Front. Psychol. 11 (329) (2020) http://dx.doi.org/10.3389/fpsyg.2020.00329.

[70] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, Rana El Kaliouby, AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, in: CHI EA '16, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450340823, 2016, pp. 3723–3726, http://dx.doi.org/10.1145/2851581.2890247.

[71] Paul Boersma, Praat, a system for doing phonetics by computer, Glot Int. 5 (9/10) (2001) 341–345.

[72] Florian Eyben, Martin Wöllmer, Björn. Schuller, openSMILE-the munich versatile and fast open-source audio feature extractor, in: Proceedings of ACM Multimedia, 2010, pp. 1459–1462, Florence, Italy.

[73] Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller, Recent developments in openSMILE, the Munich open-source multimedia feature extractor, in: MM '13: Proceedings of the 21st ACM International Conference on Multimedia, 2013, pp. 835–838, http://dx.doi.org/10.1145/2502081.2502224.

[74] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, Khiet P. Truong, The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2) (2016) 190–202, http://dx.doi.org/10.1109/TAFFC.2015.2457417.

[75] Florian Eyben, Martin Wöllmer, Björn Schuller, OpenEAR — Introducing the munich open-source emotion and affect recognition toolkit, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, pp. 1–6, http://dx.doi.org/10.1109/ACII.2009.5349350.

[76] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, Terrence J. Sejnowski, Classifying facial actions, IEEE Trans. Pattern Anal. Mach. Intell. 21 (10) (1999) 974–989, http://dx.doi.org/10.1109/34.799905.

[77] Marian Stewart Bartlett, G.C. Littlewort, Mark G. Frank, Claudia Lainscsek, Ian R. Fasel, Javier R. Movellan, Recognizing facial expression: Machine learning and application to spontaneous behavior, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, CVPR'05, 2005, pp. 568–573.

[78] Rui Sun, Elliot Moore, Investigating glottal parameters and teager energy operators in emotion recognition, in: International Conference on Affective Computing and Intelligent Interaction, Springer-Verlag Berlin Heidelberg, 2011, pp. 425–434.

[79] Dimitrios Galanis, Sotiris Karabetsos, Maria Koutsombogera, Harris Papageorgiou, Anna Esposito, Maria-Teresa Riviello, Classification of emotional speech units in call centre interactions, in: 2013 IEEE 4th International Conference on Cognitive Infocommunications, CogInfoCom, 2013, pp. 403–406, http://dx.doi.org/10.1109/CogInfoCom.2013.6719279.

[80] Jingjie Yan, Wenming Zheng, Qinyu Xu, Guanming Lu, Haibo Li, Bei Wang, Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech, IEEE Trans. Multimed. 18 (7) (2016) 1319–1329, http://dx.doi.org/10.1109/TMM.2016.2557721.

[81] Yong Ma, Heiko Drewes, Andreas Butz, Fake moods: Can users trick an emotion-aware VoiceBot? in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, in: CHI EA '21, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450380959, 2021, pp. 1–4, http://dx.doi.org/10.1145/3411763.3451744.

[82] William Alexander IV Thompson, Creating Musical Scores Inspired by the Intersection of Human Speech and Music Through Model Based Cross Synthesis (Ph.D. thesis), School of Music, Louisiana State University, 2022.

[83] Anna-Marie Ortloff, Lydia Güntner, Maximiliane Windl, Thomas Schmidt, Martin Kocur, Christian Wolff, SentiBooks: Enhancing audiobooks via affective computing and smart light bulbs, in: Proceedings of Mensch Und Computer 2019, MuC '19, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450371988, 2019, pp. 863–866, http://dx.doi.org/10.1145/3340764.3345368.

[84] Francesca D'Errico, Isabella Poggi, Tracking a leader's humility and its emotions from body, face and voice, Web Intell. 17 (1) (2019) 63–74.

[85] Thomas Schmidt, Miriam Schlindwein, Katharina Lichtner, Christian Wolff, Investigating the relationship between emotion recognition software and usability metrics, I-Com 19 (2) (2020) 139–151, http://dx.doi.org/10.1515/icom-2020-0009.

[86] Agnese Salutari, Laura Tarantino, Giovanni De Gasperis, BlocksBot: Towards an empathic robot offering multi-modal emotion detection based on a distributed hybrid system, in: Masaaki Kurosu (Ed.), Human-Computer Interaction. Technological Innovation, Springer International Publishing, Cham, ISBN: 978-3-031-05409-9, 2022, pp. 625–638.

[87] Woan-Shiuan Chien, Huang-Cheng Chou, Chi-Chun Lee, Self-assessed emotion classification from acoustic and physiological features within small-group conversation, in: Companion Publication of the 2021 International Conference on Multimodal Interaction, 2021, pp. 230–239.

[88] Gabriel Elías Chanchí Golondrino, Manuel Alejandro Ospina Alarcon, Luz Marina Sierra Martínez, Application of affective computing in the analysis of advertising jingles in the political context, Int. J. Adv. Comput. Sci. Appl. 13 (4) (2022) 554–561.

[89] Gabriel Elías Chanchí Golondrino, Luz Marina Sierra Martínez, Luz Marina Sierra Martínez, Application of affective computing in the analysis of emotions of educational content for the prevention of COVID-19, Int. J. Eng. Appl. 10 (3) (2022) 209–219.

[90] John R. Smith, Dhiraj Joshi, Benoit Huet, Winston Hsu, Jozef Cota, Harnessing ZI for augmenting creativity: Application to movie trailer creation, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1799–1808.

[91] Nicole Novielli, Daniela Grassi, Filippo Lanubile, Alexander Serebrenik, Sensor-based emotion recognition in software development: Facial expressions as gold standard, in: 2022 10th International Conference on Affective Computing and Intelligent Interaction, ACII, 2022, pp. 1–8, http://dx.doi.org/10.1109/ACII55700.2022.9953808.

[92] Shu Liu, Kevin Koch, Zimu Zhou, Simon Föll, Xiaoxi He, Tina Menke, Elgar Fleisch, Felix Wortmann, The empathetic car: Exploring emotion inference via driver behaviour and traffic context, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5 (3) (2021) http://dx.doi.org/10.1145/3478078.

[93] Zhanna Sarsenbayeva, Gabriele Marini, Niels van Berkel, Chu Luo, Weiwei Jiang, Kangning Yang, Greg Wadley, Tilman Dingler, Vassilis Kostakos, Jorge Goncalves, Does smartphone use drive our emotions or vice versa? A causal analysis, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, (36) Association for Computing Machinery, New York, NY, USA, ISBN: 9781450367080, 2020, pp. 1–15, http://dx.doi.org/10.1145/3313831.3376163.

[94] Jinxuan (Janice) Zhou, Vrushank Phadnis, Alison Olechowski, Analysis of designer emotions in collaborative and traditional computer-aided design, J. Mech. Des. (ISSN: 1050-0472) 143 (2) (2020) 021401–1–021401–10, http://dx.doi.org/10.1115/1.4047685, arXiv:https://asmedigitalcollection.asme.org/mechanicaldesign/article-pdf/143/2/021401/6558698/md_143_2_021401.pdf, 021401.

[95] Torsten Hammann, Manuel M. Schwartze, Peter Zentel, Anna Schlomann, Christiane Even, Hans-Werner Wahl, Christian Rietz, The challenge of emotions – An experimental approach to assess the emotional competence of people with intellectual disabilities, Disabilities (ISSN: 2673-7272) 2 (4) (2022) 611–625, http://dx.doi.org/10.3390/disabilities2040044, URL https://www.mdpi.com/2673-7272/2/4/44.

[96] Jose Maria Garcia-Garcia, María del Mar Cabañero, Victor M.R. Penichet, María D. Lozano, EmoTEA: Teaching children with autism spectrum disorder to identify and express emotions, in: Proceedings of the XX International Conference on Human Computer Interaction, in: Interacción '19, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450371766, 2019, pp. 36.1–36.8, http://dx.doi.org/10.1145/3335595.3335639.

[97] Akansha Singh, Surbhi Dewan, AutisMitr: Emotion recognition assistive tool for autistic children, Open Comput. Sci. 10 (1) (2020) 259–269, http://dx.doi.org/10.1515/comp-2020-0006.

[98] Sanghoon Park, Jeeheon Ryu, Exploring preservice teachers' emotional experiences in an immersive virtual teaching simulation through facial expression recognition, Int. J. Hum.–Comput. Interact. 35 (6) (2019) 521–533, http://dx.doi.org/10.1080/10447318.2018.1469710.

[99] Alina Schmitz-Hübsch, Sven Fuchs, Challenges and prospects of emotional state diagnosis in command and control environments, in: Dylan D. Schmorrow, Cali M. Fidopiastis (Eds.), Augmented Cognition. Theoretical and Technological Approaches, Springer International Publishing, Cham, ISBN: 978-3-030-50353-6, 2020, pp. 64–75.

[100] Matthew Moreno, Earl Woodruff, Exploring the effects of background music on real-time emotional expressions, performance, and confusion mediation in middle school students, Technol. Knowl. Learn. (2021) http://dx.doi.org/10.1007/s10758-021-09556-9, Online first.

[101] Blakely C. Davis, Benjamin J. Warnick, Aaron H. Anglin, Thomas H. Allison, Gender and counterstereotypical facial expressions of emotion in crowdfunded microlending, Entrepreneurship Theory Pract. 45 (6) (2021) 1339–1365, http://dx.doi.org/10.1177/10422587211029770.

[102] Dominic A. Trevisan, Marleis Bowering, Elina Birmingham, Alexithymia, but not autism spectrum disorder, may be related to the production of emotional facial expressions, Mol. Autism 7 (46) (2016) 46.1–46.12.

[103] Tina Gupta, Claudia M. Haase, Gregory P. Strauss, Alex S. Cohen, Vijay A. Mittal, Alterations in facial expressivity in youth at clinical high-risk for psychosis, J. Abnormal Psychol. 128 (4) (2019) 341–351, http://dx.doi.org/10.1037/abn0000413.

[104] Tina Gupta, Claudia M. Haase, Gregory P. Strauss, Alex S. Cohen, Jordyn R. Ricard, Vijay A. Mittal, Alterations in facial expressions of emotion: Determining the promise of ultrathin slicing approaches and comparing human and automated coding methods in psychosis risk, Emotion 22 (4) (2022) 714–724.
[105] Noldus, FaceReader: Tool for automatic analysis of facial expression: Version 6.0., 2014, Wageningen, the Netherlands: Noldus Information Technology B.V..
[106] Sebastian Fischer, Yannick Diehm, Miguel I. Dorante, Dimitra Kotsougiani, Maximilian Kueckelhaus, Muayyad Alhefzi, Ericka M. Bueno, Bohdan Pomahac, Software-based video analysis of functional outcomes of face transplantation, Microsurgery 39 (1) (2019) 53–61, http://dx.doi.org/10.1002/micr.30360, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/micr.30360.
[107] Madiha Anjum, Emotion recognition from speech for an interactive robot agent, in: 2019 IEEE/SICE International Symposium on System Integration, SII, 2019, pp. 363–368, http://dx.doi.org/10.1109/SII.2019.8700376.
[108] Turgut Özseven, Muharrem Düğenci, SPeech ACoustic (SPAC): A novel tool for speech feature extraction and classification, Appl. Acoust. (ISSN: 0003-682X) 136 (2018) 1–8, http://dx.doi.org/10.1016/j.apacoust.2018.02.009, URL https://www.sciencedirect.com/science/article/pii/S0003682X18300070.
[109] Damien Dupré, Eva G. Krumhuber, Dennis Küster, Gary J. McKeown, A performance comparison of eight commercially available automatic classifiers for facial affect recognition, PLoS One 15 (4) (2020) e0231968, http://dx.doi.org/10.1371/journal.pone.0231968.
[110] Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, Jorge Goncalves, Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets, Vis. Comput. (ISSN: 1432-2315) 37 (6) (2021) 1447–1466, http://dx.doi.org/10.1007/s00371-020-01881-x.
[111] Arne Bernin, Larissa Müller, Sobin Ghose, Kai von Luck, Christos Grecos, Qi Wang, Florian Vogt, Towards more robust automatic facial expression recognition in smart environments, in: Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments, PETRA '17, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450352277, 2017, pp. 37–44, http://dx.doi.org/10.1145/3056540.3056546.
[112] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, Marian Bartlett, The computer expression recognition toolbox (CERT), in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 298–305.
[113] Carl Vogel, Maria Koutsombogera, Rachel Costello, Analyzing Likert scale inter-annotator disagreement, in: Anna Esposito, M. Faundez-Zanuy, F. Morabito, E. Pasero (Eds.), Neural Approaches To Dynamics of Signal Exchanges. Smart Innovation, Systems and Technologies, Springer, Singapore, ISBN: 978-981-13-8949-8, 2019, pp. 383–393, http://dx.doi.org/10.1007/978-981-13-8950-4_34.
[114] Carmen Fernández-Martínez, Alberto Fernández, AI and recruiting software: Ethical and legal implications, Paladyn, J. Behav. Robot. 11 (1) (2020) 199–216, http://dx.doi.org/10.1515/pjbr-2020-0030.
[115] Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, Terrence J. Sejnowski, Measuring facial expressions by computer image analysis, Psychophysiology 36 (2) (1999) 253–263, http://dx.doi.org/10.1017/s0048577299971664.
[116] Marian Stewart Bartlett, Gwen Littlewort, Mark G. Frank, Claudia Lainscsek, Ian R. Fasel, Javier R. Movellan, Automatic recognition of facial actions in spontaneous expressions, J. Multimed. (ISSN: 1796-2048) 1 (6) (2006) 22–35.
[117] Marian Stewart Bartlett, Javier R. Movellan, Terrence J. Sejnowski, Face recognition by independent component analysis, IEEE Trans. Neural Netw. (ISSN: 1045-9227) 13 (6) (2002) 1450–1464.
[118] Jeffrey F. Cohn, Adena J. Zlochower, James Lien, Takeo Kanade, Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding, Psychophysiology (ISSN: 0048-5772) 36 (1) (1999) 35–43.
[119] Takeo Kanade, Jeffrey F. Cohn, Yingli Tian, Comprehensive database for facial expression analysis, in: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, IEEE, ISBN: 0769505805, 2000, pp. 46–53.
[120] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, Iain Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, ISBN: 1424470307, 2010, pp. 94–101.
[121] Rosalind W. Picard, Elias Vyzas, Jennifer Healey, Toward machine emotional intelligence: Analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. (ISSN: 0162-8828) 23 (10) (2001) 1175–1191.
[122] Daniel McDuff, Rana El Kaliouby, Rosalind W. Picard, Crowdsourcing facial responses to online videos, in: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, IEEE, ISBN: 1479999539, 2015, pp. 512–518.
[123] Josh P. Davis, Tim Valentine, Robert E. Davis, Computer assisted photoanthropometric analyses of full-face and profile facial images, Forensic Sci. Int. (ISSN: 0379-0738) 200 (1–3) (2010) 165–176.
[124] Lilli Cooper, Ash Mosahebi, Mark Henley, Ankur Pandya, Michael Cadier, Nigel Mercer, Charles Nduka, Developing procedure-specific consent forms in plastic surgery: Lessons learnt, J. Plast. Reconstruct. Aesthetic Surg. (ISSN: 1748-6815) 70 (3) (2017) 428–430.
[125] Hind A. Alrubaish, Rachid Zagrouba, The effects of facial expressions on face biometric system's reliability, Information (ISSN: 2078-2489) 11 (10) (2020) 485, http://dx.doi.org/10.3390/info11100485.
[126] Jianping Hao, Cross-Racial Studies of Human Vocal Tract Dimensions and Formant Structures (Ph.D. thesis), Ohio University, 2002.