INVITED PAPER

An Engineering View on Emotions and Speech: From Analysis and Predictive Models to Responsible Human-Centered Applications

By Chi-Chun Lee, *Senior Member IEEE*, Theodora Chaspari, *Member IEEE*, Emily Mower Provost, *Senior Member IEEE*, and Shrikanth S. Narayanan, *Fellow IEEE*

ABSTRACT | The substantial growth of Internet-of-Things technology and the ubiquity of smartphone devices has increased the public and industry focus on speech emotion recognition (SER) technologies. Yet, conceptual, technical, and societal challenges restrict the wide adoption of these technologies in various domains, including, healthcare, and education. These challenges are amplified when automated emotion recognition systems are called to function "in-the-wild" due to the inherent complexity and subjectivity of human emotion, the difficulty of obtaining reliable labels at high temporal resolution, and the diverse contextual and environmental factors that confound the expression of emotion in real life. In addition, societal and ethical challenges hamper the wide acceptance and adoption of these technologies, with the public raising questions about user privacy, fairness, and explainability. This article briefly reviews the history of affective speech processing, provides an overview of current state-of-the-art approaches to SER, and discusses algorithmic approaches to render these technologies accessible to all, maximizing their benefits and leading to responsible human-centered computing applications.

KEYWORDS | Affect; deep learning; emotion; ethics; prosody; real-life monitoring; responsible design; speech analysis.

Manuscript received 28 October 2022; revised 15 March 2023 and 25 April 2023; accepted 3 May 2023. The work of Chi-Chun Lee was supported by the National Science and Technology Counsel (NSTC) Taiwan under Grant 110-2221-E-007-067-MY3. The work of Theodora Chaspari was supported by the National Science Foundation (NSF) under Grant IIS-2046118. The work of Emily Mower Provost was supported by NSF under Grant RI-2006618. The work of Shrikanth S. Narayanan was supported by NSF under Grant IIS-2204942.
(Chi-Chun Lee and Theodora Chaspari contributed equally to this work.)
(Corresponding author: Theodora Chaspari.)

Chi-Chun Lee is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan (e-mail: cclee@ee.nthu.edu.tw).
Theodora Chaspari is with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: chaspari@tamu.edu).
Emily Mower Provost is with the Department of Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: emilykmp@umich.edu).
Shrikanth S. Narayanan is with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90007 USA (e-mail: shri@ee.usc.edu).

Digital Object Identifier 10.1109/JPROC.2023.3276209

I. INTRODUCTION

Affective technologies enable computers to identify, process, and simulate human affect. They lie at the intersection of engineering, computer science, psychology, and cognitive science. Human affect experiences and displays

may vary in several ways, including their duration, intensity, and specificity, and can play an important role in regulating human cognition, psychology, and interpersonal interactions. Since human affect is inherently multimodal, affective computing technologies typically rely on observable signals, such as facial expressions, body language, and/or speech characteristics.

The focus of this article is on affective speech technologies—technologies that process and characterize speech signals to identify experienced and expressed affect. We briefly review the history of affective speech processing and provide evidence from speech science that explains the modulation of affect in speech. Following that, we outline knowledge-driven approaches to the quantitative analysis of speech affect, as well as more recent work on data-driven affect representations. Motivated by the increasing prevalence of ambulatory speech technologies, we discuss applications of affective speech technologies in domains such as healthcare and education, and the associated technical challenges. These include building generalizable models and personalizing the models to individuals, for example, to enhance their technology-mediated user experiences. Finally, we outline sociotechnical challenges in affective speech technologies, which includes preserving privacy, being inclusive, ensuring equitable outcomes, and promoting transparency through the explainability of system behavior for improving human–AI interaction in decision-making. We also discuss ethical issues that arise from the wide use of these technologies outside the laboratory in real-world settings (see Fig. 1).

II. BRIEF HISTORY

Psychological sciences have proposed various conceptual frameworks for modeling affect. Affect refers to a broad range of experiences that people can have and embodies both emotions and moods. Emotion is a short-term, often intense, experience that is typically directed at a source [1] and is only part of affect. Emotion representation usually relies on either a discrete representation of (basic) emotions (e.g., anger, fear, and happiness) or a continuous representation of arousal/valence/dominance [2]. In contrast to emotion, mood is a state of mind that tends to be less intense than an emotion and does not necessarily need a stimulus. Overall, moods last longer than emotions. Mood can be measured with psychologically validated scales that are designed in the context of general applications, such as the Pick-A-Mood scale [3], and clinical applications [4], [5], such as the Hamilton Depression Scale (HamD) [6] and the Young Mania Rating Scale (YMRS) [7].

Different emotions are characterized by unique speech patterns [8]. Scientific studies on emotional speech production have focused on the premise that a speaker's emotional state influences the neuromuscular control of vocal organs—both the voice source and the supralaryngeal articulators, such as the tongue, jaw, and the lips—and the resulting aeroacoustic mechanisms and (signal) consequences. Many classic studies have analyzed the acoustic characteristics, in particular, the expressed prosodic and spectral aspects of emotional speech [9], [10], [11]. Speech patterns of pitch and amplitude modulation and segmental duration (prosody) carry affective information [11]. For example, fundamental frequency (F0) tends to increase in highly activated emotional speech compared to neutral speech [12]. Thus, we can quantify speaker affect by encoding certain characteristics that can be amenable to objective measurement in the speech signal.

Fewer studies have analyzed direct articulatory details of emotional speech production. Studies using flashpoint tracking using electromagnetic articulography have shown more peripheral or advanced tongue positions than neutral speech articulation and have demonstrated that the movement range of the jaw is larger for angry speech compared to neutral, sad, or happy speech [13], [14], [15]. Prior work has also revealed the relationship between the variability of an articulator and the linguistic criticality of the articulator in emotional speech [16]. A more complete view of the dynamic vocal tract afforded by real-time magnetic resonance imaging has enabled further detailed analysis of speech-emotional articulation that reveals both speaker-dependent and speaker-independent variation patterns [17]. For example, sad speech, a low arousal emotion, tends to show a smaller opening for low vowels in the front cavity than the high arousal emotions more consistently than the other regions of the vocal tract. Happiness depicts significantly shorter vocal tract length than anger and sadness in most speakers. Together, these acoustic and articulatory studies underscore the inherent variability expected in emotional speech expressions within and across [18], [19], which needs to be contended with by computational methods and models.

Scientific findings from the field of speech production have informed engineering approaches for quantifying emotions expressed in speech. F0, which is a measure of pitch (i.e., the vibration frequency of the vocal folds), is a widely used measure that tends to depict differences between emotional speech and neutral speech [12]. Anger and happiness depict increased F0 compared to neutral emotions [20], but these findings are confounded by the emotion elicitation method, recording conditions, and linguistic content of an utterance. The variation of F0 over linguistic categories, such as accents (i.e., syllables with prominence) [21], vowels and consonants [22], [23], or stressed and unstressed syllables [22], has been also extensively examined. For example, happiness and anger depict the largest F0 range over sentence and word accents, while fearful and neutral emotions depict the lowest F0 range [21]. Other work has also demonstrated the presence of emotional content in various parts of the pitch contour, such as the end (i.e., final rise), or the relevance of the direction of the pitch contour in conveying emotion [24]. Additional speech characteristics, such as voice level, voice quality, articulation precision, and spectral parameters, have been investigated in the context

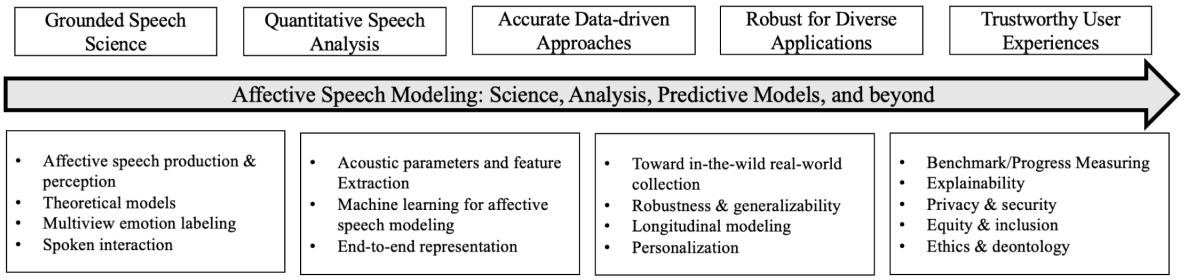


Fig. 1. ***Schematic overview of current and future speech emotion technologies: transitioning from speech science, quantitative analysis, and representation learning to trustworthy real-life user experiences.***

of emotional speech [24]. Grounded in these findings, a large set of studies on emotional speech synthesis have attempted to encode emotion in a speech via controlling the prosodic parameters [25], [26].

The expression of emotions in speech sounds and one's ability to perceive those emotions from the speech are fundamental to human communication [27]. A speaker can express their emotion via modulating the acoustic properties of their speech, while a listener can potentially perceive this modulation of acoustic properties and infer the emotion of the speaker. Prior studies have conducted perceptual experiments in which listeners were asked to identify the perceived emotions in speech signals with results indicating that listeners were able to identify the intended emotions at rates significantly higher than chance levels [28], but still there is a mismatch between what was intended and what was perceived [29]. This mismatch can become more pronounced in speakers that suffer from clinical conditions, such as neurological disorders [30], [31]. Research has also explored the ability to infer emotion that transcends different languages and cultures [32] with evidence both in favor and against this proposition [33]. While some researchers view emotion as a universal construct given the biological basis of the emotional experience, wide evidence supports the presence of cultural differences in emotion expression and perception [34], [35]. Given the fundamental role of emotion in human communication, the expression and perception of emotion have been the matter of investigation in diverse domains of human interaction, such as adult–child interaction, medical encounters, clinical therapy, and job interviews [36].

III. CURRENT STATE

This section reviews the current state of speech emotion recognition (SER) technologies. It first discusses the collection, curation, and annotation of emotion speech corpora (see Section III-A). It further outlines technical approaches to modeling emotions from speech, including feature design and representation learning methods (see Section III-B).

A. Construction of Affective Speech Corpora: From in-Lab to in-the-Wild Datasets

Datasets with clear and varied emotional content are generally either *directly collected* [38], [44], [45], [60] or *curated* based on existing digital resources [47], [52], as summarized in Table 1. Prior review studies have provided a detailed overview of existing datasets built for the task of SER that includes a speech from various languages [61], [62]. The benefits of direct collection methods follow from the controlled nature of the environment, the clearly defined elicitation protocols, and even the predefined lexical content. These controls provide an opportunity to highlight the (expressions of) emotionally relevant content while mitigating the effects of unwanted variability. However, the reality is that, in deployed environments, unwanted and confounding variability is pervasive. This has led researchers to embrace curation methods. In these methods, the data have no centralized collection location or platform. However, the data curated are generally performative (i.e., either podcasts [47] or YouTube videos [52]), rather than from natural dialogs, again raising the specter of mismatch with respect to an intended use case. There exist datasets that are collected in natural contexts, for example, those focused on the overlap between mental health and emotion [63]. However, due to privacy concerns, datasets of this type are rarely publicly available.

Therefore, there is a mismatch that exists between publicly available datasets and natural conversational use cases. This mismatch forces us to rethink the interplay between algorithm development and data collection, and could lead us to advocate for truly natural data collection protocols. However, the practices of data collection are themselves a reaction to the challenges in obtaining truly representative data. If we move into real-world recording environments, in which individuals interact naturally and without constraints, we must be prepared to answer a new set of questions: 1) how do we record data in the real world from a practical and ethical standpoint? 2) how do we obtain annotations for data that may be infused with both contextual and cultural variability? and 3) how do we measure progress as a field?

An alternative to the above is to rethink how we annotate and use publicly available collections of data. These collections contain not only emotionally expressive data but also environmental data. Emotional data can be augmented with environmental data to expose SER classifiers to ambient conditions that are more likely in real-world environments [64], [65], [66] and allow for

Table 1 Overview of Most Commonly Used Speech Datasets of Emotion

Corpus	Year	Language	Speaker No.	Elicitation Method	Data Amount (hrs)	Emotion Attributes
eNterface05 [37]	2006	English	42	Acted	2.58	6 classes
IEMOCAP [38]	2008	English	10	Acted	12	9 classes
SAVEE [39]	2010	British English	4	Induced	3.84	7 classes
RECOLA [40]	2013	French	46	Induced	9.5	Valence/Arousal
AVEC [41]	2014	German	292	Induced	240	Valence/Arousal/ Dominance
CREMA-D [42]	2014	English	91	Acted	5.3	6 classes
BAUM-2 [43]	2015	English/Turkish	286	Natural	1	7 classes
MSP-IMPROV [44]	2016	English	12	Acted	3	10 emotions; Valence/Arousal/ Dominance/Naturalness
CreativeIT [45]	2016	English	16	Acted	8	Valence/Arousal
emoFBVP [46]	2016	English	10	Induced	-	23 classes
MSP-PODCAST [47]	2017	English	1458	Natural	166	16 classes Valence/Arousal/Dominance
NNIME [48]	2017	Mandarin Chinese	43	Acted	11	10 classes Valence/Arousal
CHEAVD2.0 [49]	2018	Chinese	527	Induced	8	8 classes
RAMAS [50]	2018	Russian	10	Induced	7	6 classes Domination/Submission
MELD [51]	2018	English	407	Induce	13	7 classes
CMU-MOSEI [52]	2018	English	1000	Natural	66	6 classes; Sentiment
SUBESCO [53]	2019	Standard Bangla	20	Acted	7.7	7 classes
SEWA [54]	2019	Chinese/English/German Greek/Hungarian/Serbian	398	Induced/ Natural	44	Valence/Arousal
CH-SIMS [55]	2020	Mandarin	474	Natural	2.5	5 classes
IESC-Child [56]	2020	Spanish	174	Induced	34.9	8 classes; Valence
DEMoS [57]	2020	Italian	68	Induced	7.5	7 classes; Valence/Arousal
MSP-Conversation [58]	2020	English	197	Natural	15.15	Valence/Arousal/Dominance
M3ED [59]	2022	Mandarin	626	Natural	50	7 classes

data encoding differences [67]. Commonly used examples include the Dataset for Environmental Sound Classification (ESC) [68], Audio Set [69], and DEMAND [70]. A similar aim can be accomplished by signal manipulations, such as speeding up the audio (e.g., with pyAudio [71]) or introducing reverberation effects (e.g., with Pedalboard [72]). However, even when datasets that enable augmentation to promote robustness to environmental differences are available, models constructed based on this data may still need to handle variations due to differences in the recording devices themselves (e.g., different channel characteristics of the microphones [5]). Furthermore, the methods designed to augment datasets may themselves change how evaluators interpret the emotional information [73].

Beyond the collection of audio data, the inherent ambiguity in the signals expressing emotion makes it an open question as how to obtain reliable labels to support affective speech technologies. Affect annotations can be conducted in a holistic (i.e., characterizing the overall emotion content of an utterance) or continuous (i.e., characterizing emotion over a continuous time scale) manner. Acted speech data are typically assigned to the predefined emotion that was originally intended during the elicitation method. In order to verify the reliability of the emotional content, prior work has proposed to obtain perceptual ratings for each utterance and consider as valid only the utterances that depict high interannotator agreement [42]. Evoked and natural emotion elicitation methods further require human annotations that can be obtained either via self-assessment or third-party labels. The annotation of affect is an inherently complex task that is confounded by the unique experiential, cultural, and technical background of the one who is called to perceive emotion [74], especially when it comes to continuous ratings [75], [76].

Self-assessment provides one's perspective about their emotion. It is useful and widely adopted but many times suffers from subjectivity and recall bias [77] and may differ from third-party labels [78]. One common type of self-assessment is ecological momentary assessment (EMA) [79], [80], which occurs within the data collection period (e.g., usually a few times per day when obtaining longitudinal data) and, thus, attempts to minimize recall bias and maximize ecological validity. Third-party labels give a spectator's perspective about the target's emotion and are usually obtained by experts or via crowdsourcing. These focus on the expression of emotion that is observable and are confounded by the reliability or expertise of annotators, the temporal segmentation of the data, knowledge of context, and the potential temporal misalignment among annotators. When obtaining labels via third-party annotations or self-reports, one should also be mindful of multicultural differences in decoding emotions since individuals tend to resort to stereotype knowledge and norms that are prone to their sociocultural background when decoding subtle emotional expressions [81]. A metareview on culturally specific elements of emotion perception can be found in [74].

The study of interannotator agreement in third-party annotations of affect has been another important focal point of the research community. Existing work provides evidence of an overall moderate interannotator agreement when rating affect [82], [83], which can be mitigated by various "good practices," such as recruiting annotators with motivation and prior experience, constructing a well-defined annotation manual, providing annotators with an overall view of the content

that they will be assigned to, and learning/integrating annotator-specific distortions in the data analysis [75], [84]. Several approaches have been proposed to model annotator-specific distortions that might yield from differences in the reaction time and emotional perception. Mariooryad and Busso [85] compensated for differences in reaction time among annotators by estimating an annotator-specific time shift via maximizing the mutual information between the signal-based behaviors and the time-continuous annotations. Following that, Gupta et al. [84] investigated a comprehensive model of annotator-specific distortions in both space and time that were approximated as a linear filter. Trigeorgis et al. [86] further examined a method that temporally aligns the annotations with embeddings learned from the data, which was also compared against conventional time alignment approaches, such as the dynamic time wrapping (DTW) and the canonical correlation analysis (CCA). Khorram et al. [87] introduced a convolutional neural network (CNN) that simultaneously aligns and predicts labels in an end-to-end manner. Ramakrishna et al. [88] proposed a multidimensional annotation fusion that jointly models the various affective dimensions leading to more accurate ground truth estimates (again treated as latent variable distorted by the annotators) and is applicable to both global and time series annotation fusion problems.

As new datasets and new annotation schemes come online, there are questions that arise surrounding metric design and progress tracking. There are two predominant metrics in the field: unweighted average recall (UAR) and concordance correlation coefficient (CCC) for discrete and continuous outcomes, respectively. The predominance of these two metrics, and of the common emotion recognition datasets, provides an opportunity to create a central repository for results. This will allow the field to have a straightforward method for tracking progress and evaluating newly proposed approaches. However, the challenge with this centralization lies in the assumption that these datasets and metrics are the most appropriate for the field. Available datasets mostly include utterances from high-resource languages, while it would be important to establish benchmarks for low-resource languages as well. In addition, evaluation metrics, such as UAR and CCC, were constructed with the assumption of a deterministic label (i.e., either discrete or continuous). As we move to more natural and ambiguous expressions of emotion, the relevance of a single label may become overly restrictive, which, in turn, may necessitate the creation of new metrics or new styles of evaluation (e.g., soft labels [89], [90]). In addition, with the advent of new social challenges, it would be important to consider other metrics, such as the inclusiveness of sociodemographic and linguistic characteristics of speakers in the dataset and the consideration of evaluation metrics beyond performance, such as fairness and explainability.

Alternative annotation schema may also address emotional ambiguity. We describe emotion with words, so it is natural to expand the vocabulary to cover more specific emotions. For example, Lotfian and Busso [47] employed a secondary set of eight emotions in addition to the primary emotion words to encode emotional ambiguity in naturalistic speech, but much larger sets are sensible depending on the use case. Associated challenges in constructing the word set include how annotators and end-users perceive words, how to size the set while considering redundancy and specificity, and which types of words are relevant to the expressed emotion. Challenges with annotation include metrics for both modeling and annotation, ways to handle label sparsity, how to encode complex expressions along with annotator uncertainty, and what annotators attune to; for example, further investigation needs to be conducted on whether annotators may focus on overt prosody, intent, or suspected internal emotion—each requiring progressively more context.

B. Modeling of Emotion From Speech: From Feature Design to Deep Learning Representations

Early methods of modeling emotion relied on the design of knowledge-driven features that capture acoustic patterns of emotion modulation in speech. Affect conveyed in the voice has been empirically documented by the measurement of parameters of phonation and articulation [91], [92]. Examples include parameters in the time domain (e.g., speech rate), the frequency domain (e.g., F0 and formant frequencies), the amplitude domain (e.g., intensity or energy), and the spectral distribution (e.g., Mel-frequency cepstral coefficients (MFCCs) and relative energy in different frequency bands). These descriptors are typically combined with statistical measures to summarize temporal trajectories into a fixed-sized vector. In an early exploratory work, Lee et al. [93] demonstrated that emotion patterns are the most prominent in vowel sounds, whose spectral features yielded increased emotion classification performance compared to the spectral features computed on the entire speech signal. Bone et al. [94] proposed an unsupervised rule-based framework using knowledge-driven acoustic features to quantify the level of arousal in an utterance. More recently, a group of voice and speech scientists convened at an interdisciplinary meeting in Geneva and devised an effective subset of voice parameters with theoretical significance, named “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS).” Features included in GeMAPS capture the speech energy content and its relative change across frequency bands that are theoretically stipulated and practically demonstrated to relate to emotion [95]. In order to promote reproducible research, many of the above efforts have been publicly disseminated in open-source toolboxes, such as the openSMILE [96] and COVAREP [97], yielding off-the-shelf acoustic descriptors of emotion that can be readily used.

The field of SER has progressed tremendously with the advent of deep learning methods since learning discriminative representations is at the core of advancing

the observed performances. In one of the early works in this domain, Mao et al. [98] proposed the learning of affective speech directly from the speech spectrogram. The authors leveraged a variant of a sparse autoencoder with reconstruction penalization to learn local invariant features from all the unlabeled samples. The learned features further served as the input to a discriminative learning task with a loss function that encouraged saliency, orthogonality, and discrimination for emotion recognition. Following that, Trigeorgis et al. [99] proposed an end-to-end learning framework of emotion classification from raw speech. Raw speech segments of 6-s length served as the input to a set of 1-D convolution and max-pooling operations. The temporal evolution within each segment was further modeled via a general-purpose bidirectional long short-term memory (LSTM) network. After training the LSTM, the authors found that the corresponding cell activation was associated with prosodic features (e.g., loudness and F0). Other approaches have demonstrated that integrating temporal context in the short term (i.e., modeling the multitemporal evolution of frame-level representations) and the long term (i.e., aggregating those learned features into a compact utterance-level representation) can benefit emotion classification [100], [101].

Benchmarking SER deep learning algorithms is naturally challenging due to the aforementioned complexity in collecting data and labels (see Section III-A). In order to illustrate key components of SER research, we list as an example those SER works conducted on the IEMOCAP, MSP-Podcast, and MSP-IMPROV, which are commonly used in previous deep learning studies (see Table 2). We provide a list of key research with the corresponding accuracy that is obtained across various settings, including the choice of machine learning types (e.g., supervised, semisupervised, and multitask), input speech features (e.g., Wav2Vec [102], raw waveform, and log-mel spectrogram), target labels, and used evaluation metrics. Although this summary is not exhaustive, it provides a glimpse into the states and accuracy numbers obtained for modern SER research. Note that the majority of state-of-the-art results are obtained by utilizing variants of transformer-based architectures of pretrained speech representations (e.g., Wav2Vec [102] and HuBERT [103]). Interested readers can refer to a more comprehensive review of deep representation learning for affective speech processing provided by Lee et al. [104]. Despite the promising results, deep learning methods are still susceptible to the small-scale emotion speech datasets (i.e., relatively to larger scale datasets used in conventional speech processing and computer vision tasks), the limited generalizability of the representations in other domains, and the lack of explainability, as will be discussed in Section IV.

IV. WHAT'S NEXT

The SER community has considerably matured over the last few years. Yet, in order for emotion speech technologies to truly make a positive societal impact, they need to progress so that they can be effectively and responsibly integrated into everyday lives. To achieve this, it is important to reorient the focus of the field toward three directions: 1) conducting data collection efforts that are aligned with the focal use case and devising new annotations strategies for the collected data (see Section IV-A); 2) addressing engineering challenges that hamper the generalizability of emotion recognition systems to multiple use cases while, at the same time, ensuring a personalized experience for each user (see Section III-B); and 3) identifying societal challenges related to affective speech technologies, modifying, or redesigning these technologies to address these challenges with a focus on benefiting humans, allowing users to effectively calibrate their trust in speech emotion technologies, and effectively disseminating emotion speech technologies via raising public awareness regarding their benefits while, at the same time, listening to potential public skepticism or criticism (see Section IV-C).

A. Application-Specific Data Collection, Annotation, and Modeling

One's ability to effectively manage emotions is critical to healthy psychological and social development. Thus, affective speech technologies have become increasingly popular foci in the context of mental health, including mood (e.g., depression), developmental [e.g., autism spectrum disorder (ASD)], and neurological (e.g., Parkinson's) disorders. Traditionally, the monitoring of mental health conditions is performed at low temporal resolution via third-party evaluations based on clinician and caregiver reports, in addition to self-reports. The enhanced capability and affordability of ambulatory sensing devices have led to an increasing focus on speech-based mental health monitoring algorithms [122]. If designed effectively and responsibly, these algorithms can unobtrusively track one's condition, explore its longitudinal variation in real life, and even predict the onset or episodes of mental health degradation [123]. These can potentially result in novel insights regarding mental health disorders, promote more personalized treatment, and increase the engagement of users and patients [124]. In the following, we will overview prior work on affective speech technologies for mental health by exemplifying our discussion on three different cases of health conditions, including depression, ASD, and Parkinson's disease.

Depression is a serious but common mental health condition that affects one's feelings, thoughts, and daily functioning [125]. Patients with depression have been characterized by distinct vocal prosodic patterns, such as decreased speech loudness, slowed speech rate, and monotonous pitch [126], [127]. Beyond the ease of acquisition, an additional advantage of audio-only analysis—without accessing the linguistic or other accompanying visual cues—is the increased potential for privacy. Thus, the automated assessment and tracking of markers of

Table 2 Examples of Recent Deep Learning Approaches for Speech Emotion Recognition

Author	Learning Type	Evaluation Databases	Feature Input	Classes	Performances
Latif et al., 2022 [105]	SemiSRL + MTRL	IEMOCAP/MSP-IMPROV	Spectrogram	A, H (E), N, S/A, H, N, S	UAR: 0.688/0.636
Zhang et al., 2022 [106]	ADRL	IEMOCAP/MSP-IMPROV	LogMel	A, H (E), N, S/A, H, N, S	UAR: 0.523/0.473 F1: 0.333/0.5
Chou et al., 2022 [90]	MTRL	MSP-Podcast	Wav2Vec	8-class primary 16-class secondary	F1:0.316 F1:0.38
Chou et al., 2022 [90]	SRL	MSP-Podcast	Wav2Vec	8-class primary	F1: 0.356
Li et al., 2022 [107]	SRL	IEMOCAP	Wav2Vec2/MFCC	A, H (E), N, S	ACC: 0.649
Gat et al., 2022 [108]	SelfSRL	IEMOCAP	Hubert	A, H (E), N, S	ACC: 0.742
Feng et al., 2022 [109]	SemiSRL	IEMOCAP/MSP-IMPROV	Raw Speech	A, H, N, S/A, H, N, S	UAR: 0.682/0.557
Pepino et al., 2021 [110]	SRL	IEMOCAP	Wav2Vec2	A, H, N, S	UAR: 0.672
Cai et al., 2021 [111]	MTRL	IEMOCAP	Wav2Vec2	A, H (E), N, S	UAR: 0.782
Muhammad et al., 2020 [112]	SRL	IEMOCAP	Spectrogram	A, H (E), N, S	ACC: 0.723
Latif et al., 2020 [113]	SRL	IEMOCAP	IS10	A, H (E), N, S/A, H, N, S	UAR: 0.596
Nediyanchath et al., 2020 [114]	MTRL	IEMOCAP	LogMel	A, H, N, S (improvised only)	ACC: 0.764 UAR: 0.701
Hao et al., 2019 [115]	SRL	IEMOCAP	LogMel	A, H, N, S (improvised only)	ACC: 0.692
Latif et al., 2019 [101]	SRL	IEMOCAP/MSP-IMPROV	Raw Speech	A, H (E), N, S	UAR: 0.602/0.524
Paraskevopoulos et al., 2019 [116]	URL	IEMOCAP	IS10	A, H (E), N, S	UAR:0.598
Tao et al., 2019 [117]	SemiSRL	IEMOCAP	IS09	A, H (E), N, S	UAR:0. 596
Latif et al., 2018 [118]	URL	IEMOCAP	LogMel	A, H (E), N, S	UAR: 0.581 ACC: 0.611
Eskimez et al., 2018 [119]	URL	IEMOCAP	MFCC + LLDs	A, E, FU, N, S	UAR: 0.485 F1: 0.461
Lotfian et al., 2018 [120]	MTRL	MSP-Podcast	eGeMAPs	8-class primary	F1: 0.263
Tao et al., 2018 [121]	MTRL	IEMOCAP	IS10	A, H, N, S	ACC: 0.587 F1: 0.582

SRL: Supervised Representation Learning; SemiSRL: Semi-Supervised Representation Learning; URL: Unsupervised Representation Learning
MTRL: Multi-Task Representation Learning; ADRL: Domain Adaptive Representation Learning; DRL-RL: Deep Reinforcement Learning
SelfRL denotes Self Supervised Representation Learning
A: Angry, H: Happy, E: Excited, N: Neutral, S: Sad, Fu: Frustrated
8-class primary: Anger, Sadness, Happiness, Surprise, Fear, Disgust, Contempt, Neutral
16-class primary: 8-class primary, Amusement, Frustration, Depression, Concern, Disappointment, Excitement, Confusion, Annoyance

depression from vocal acoustics have received wide interest from the speech community, exemplified in extensive survey papers [127], [128], and publicly available datasets, such as the Distress Analysis Interview Corpus (DAIC) [129] and data included in the Audio-Visual Emotion Challenge (AVEC) 2017 and 2019 challenges [130], [131]. Early work on automatic depression assessment has examined prosody, source features, formants, and spectral measures that serve as the input to machine learning classifiers for estimating the presence or severity of depression [132], [133], [134], [135]. Expanding upon this work, other studies have focused on longitudinal mood monitoring via smartphone devices in order to track depression symptom severity for individuals with bipolar disorder [5], [63], [136] and major depressive disorder [137]. Efforts have also focused on multimodal approaches that incorporate linguistic and visual information to vocal acoustics [138]. More recent work investigated end-to-end representation learning that relies on learning depression-specific patterns in the speech spectrogram. For example, DepAudioNet was introduced as an effective deep learning method that modeled short-term temporal and spectral correlations within the speech spectrogram with a CNN, followed by capturing temporal correlations across speech frames via an LSTM network [139]. Grounded in evidence that depression can affect phoneme-level variations in speech [140], AudVowelConsNet learned spectrogram-based features related to depression based on the consonant and vowel regions of speech [141]. Temporal dependencies at the frame level (or vowel level), the word level, and the sentence level were further modeled in a hierarchical manner via attention mechanisms [142], [143].

ASD is a heterogeneous condition characterized by difficulties with social communication and social interaction, and restricted and repetitive patterns in behaviors, interests, and activities [144]. A core symptom of ASD involves segmental and suprasegmental speech atypicality that primarily resides in the pragmatic and affective aspects of prosody, manifested in monotonous pitch, deficits in speech volume control and vocal quality, and atypical stress patterns [145], [146], [147]. Research on speech-based automated ASD assessment has examined prosodic measures [148], [149] and spectral parameters [150], as well as various representation learning methods [151], [152]. Another line of work has focused on the early detection of ASD for infants and toddlers via investigating spectral parameters (e.g., spectral flatness) and phonation characteristics (e.g., jitter and shimmer) [153], [154], [155]. Beyond assessment or early detection, affective computing technologies can quantify potential disruptions in affect-related processes of individuals with ASD, including the relation between voice and face in affective expression and perception in ASD under various interaction contexts [156]. Insights from this analysis can contribute to novel therapy mechanisms and interventions via identifying specific parts of an interaction that can be beneficial to a patient with ASD and tailoring therapy sessions to a patient via predicting beneficial interaction strategies on a moment-to-moment basis [157], [158].

Parkinson's disease is a neurological disorder that is characterized by uncontrollable or unintended

movement [159]. Given that speech production relies on the sophisticated movement of vocal organs, it is often impacted in patients with Parkinson's disease and revealed in speech cues such as imprecise articulation and breathy or hoarse voice quality [160], [161]. Thus, prior work has focused on classifying between patients with Parkinson's disease and their healthy counterparts based on speech intelligibility metrics [162], [163], speech dysphonia measures that quantify time- and amplitude-based speech periodicity and variations of this periodicity [164], and prosodic and vocal fold excitation parameters [165]. In addition, patients with Parkinson's disease often depict monotonous pitch and loudness, which can result in the corresponding speech being perceived as "sad," a condition also referred to as "dysathria" [166]. Prior work further provides evidence of impairment in emotional expression and recognition for patients with Parkinson's disease, who tend to be less accurate than healthy individuals in the recognition of negative emotions [31], a finding that also applies to the perception of negative emotions from voice [30]. The expressive quality of speech is also affected. Patients with Parkinson's disease have difficulty expressing emotional prosody [167], [168], a finding that has been corroborated via perceptual user studies [169] and experiments of automatic speech-based emotion recognition [170]. Given this evidence of impaired emotion perception and expression, affective speech technologies can offer methods for supporting disease screening and progression monitoring of patients with Parkinson's disease via investigating changes in emotional speech, potentially focusing on improvements in the variability of pitch and loudness after therapy and medication intake.

Beyond healthcare, affective speech technologies can be used in other domains, such as education and decision support. In the field of education, tracking student affect can contribute toward an emotionally aware and personalized learning experience that can effectively support the instructor by suggesting real-time adjustments to the class content and activities [171], [172], [173], [174]. In team science, monitoring individual and team affect can improve our understanding of specific types of behaviors that are beneficial or detrimental to team functioning and, thus, help mitigate interpersonal conflict [175], [176]. A recent line of work focuses on quantifying subtle socioemotional behaviors in teams, such as microaggressions and microaffirmations, via conversational markers, and investigating their association to team affect and performance, especially in the context of diverse teaming [177]. Given that emotions constitute pervasive drivers of decision-making [178], prior work highlights the important of modeling user affect in decision support [179]. The importance of modeling situational factors, such as affect, is becoming even more prevalent in recent decision support paradigms that involve interactions between human users and AI algorithms [180], [181], which are already confounded by various human- and system-related factors [182].

Human affect in the above settings might be complex and multifaceted, thus requiring domain-appropriate strategies for data capturing, annotating, and modeling. It is recommended that one collaborates with domain experts across all stages, from the original research design to the actual technology implementation, and in an iterative manner that allows for fine-tuning and potentially redesigning the technology. It is also necessary that one relies on existing theoretical models and conceptual foundations so that the technology is developed in tandem with the specific needs of the stakeholders. Obtaining insights into the data and affect-specific patterns is also essential. This can be achieved via analyzing knowledge-driven measures related to segmental and suprasegmental information in association with the considered outcomes and by seeking interpretable explanations from potentially complex representation learning models. Since affect does not depict similar temporal manifestations to other conditions, such as the ones related to mental health (e.g., depression is a condition that slowly changes over time, while emotions are short-term experiences), it is important that one considers appropriate temporal frameworks for the analysis of the data and the evaluation of the systems, and takes into account the interaction across potentially distinct temporal scales. Given the sensitivity and high stakes associated with the aforementioned applications, the research community as a whole should define appropriate operational thresholds in conjunction with stakeholders and policy makers. Toward this direction, it is valuable to create common benchmarks publicly available to researchers, and render data (as permitted) and algorithms readily available in public repositories.

B. Engineering Challenges in Affective Speech Technology for Real-Life Applications

1) Promoting Robustness and Generalizability: With the advent of speech technologies in various applications, it becomes more and more important to learn generalizable representations that can respond to novel situations not observed in the training data and handle cross-domain mismatch. In fact, prior work in SER has investigated various supervised, semisupervised, and unsupervised learning methods to effectively transfer learned speech emotion representations among domains, even in the presence of limited labeled data (see Table 2). In terms of supervised learning, previous studies have explored the effect of feature normalization and supervised domain adaptation, such as adaptive and incremental support vector machines. For instance, Abdelwahab and Busso [183] proposed an adversarial learning framework that is composed of an emotion classifier and a domain classifier. The domain module of the network learns a representation that confuses a competent domain classifier. This leads to models that perform better in the target domain without impacting the performance in the source domain. Gideon et al. [184] proposed a continuous domain adaptation method designed to create representations that "meet

in the middle" between disparate datasets. The domain adaptation leveraged Earth movers distance between representations from source and target domains to aid in the learning process [184]. Other approaches have proposed to leverage dependencies in speech information between interconnected attributes, such as emotion, gender, and speaker identity, working off the assumption that learning speech variations across multiple constructs can result in more generalizable representations [114], [185], [186]. Metric learning approaches that model relative differences between emotions in the learned feature embedding have been also investigated as few-shot learning approaches generalizing emotion representations with few labeled samples [187]. In a most recent effort, EmoNet was assembled from 26 SER corpora (∼65 hours data) and was built to serve as a generalizable multicorpus SER model [188]. Transfer learning methods can have a significant impact on real-life affect recognition, where labels are noisy and difficult to obtain. An interesting future direction would be to leverage the ample amount of labeled data from in-lab datasets to pretrain affect recognition models that can be readily and effectively translated into real-world situations with minimal or no supervision.

2) Personalizing the User Experience: The large-scale acquisition of speech data is redefining the delivery of affect recognition technologies. Combined with AI capabilities, affective speech technologies can now target users in a personalized manner and address their needs in the context of specific circumstances. The motivation for personalization is conceptually grounded in psychological sciences arguing large interindividual variability in the ways that emotions are experienced and expressed [18], [189]. For instance, King and Emmons [190] found that individuals who are overall more ambivalent about expressing their emotions are less likely to actually express their felt emotions, a finding that could potentially be reflected in their speech patterns. On the basis of the hypothesis that emotional dispositions are core characteristics of personality dimensions [18], other studies have explored personality differences in emotions. Indicatively, Ng and Diener [191] found that individuals who score high in neuroticism have a stronger feeling of more negative emotions compared to their counterparts, while those who score high in extraversion have a higher feeling of positive emotions. Such evidence from the field of psychology has motivated affective computing researchers to study personalized models for affect recognition. It is recommended that affective computing researchers overview studies from psychological sciences in tandem with the model development since the integration of domain-specific information might contribute to the reliability and interpretability of the final personalized models.

Recent studies have examined various approaches for effectively recognizing emotions from each user. Rahman and Busso [192] examined an unsupervised feature adaptation scheme that aims to reduce the mismatch between features corresponding to a general model (i.e., trained for all users) and the ones corresponding to a target user. Li and Lee [193] proposed a personalized attention mechanism via constructing a personal profile embedding that compares the psycholinguistic attributes of the target speaker to the other speakers from a large-scale dataset. Vryzas et al. [194] leveraged transfer learning methodologies to tailor the learning of speech-based emotion recognition models. Models were pretrained on in-domain data from all speakers or on publicly available out-of-domain data and subsequently fine-tuned on the target speaker.

Personalized models can be particularly beneficial in mood monitoring allowing the assessment and management of mood disorders. Karam et al. [195] examined the use of speech patterns associated with mood transitions, varying from a healthy euthymic state to states characterized by mania or depression, using structured (i.e., weekly clinical interactions) and unstructured (i.e., calls recording participants' speech outside the clinical interaction) speech recordings. Khorram et al. [196] demonstrated that the prediction of depression from these data could be made significantly more accurate by first personalizing the models using speaker embedding features. Arevian et al. [109] also proposed a longitudinal speech tracking system that recorded speech data from individuals with mental health conditions over the span of four months, which were used to develop a machine learning system for clinical state tracking based on acoustic and lexical measures. Yan et al. [197] leveraged multimodal data, including speech, to track daily changes in affect over the span of eight weeks. Findings indicate that subject-dependent normalization of affect labels can help improve the performance of the machine learning models, and the leveraging of previously collected data from an individual can contribute to the reliable estimation of affect in future time points.

C. Responsible Affective Speech Technology

Intelligent ambulatory technologies have witnessed public backlash and wide skepticism due to the ethical and societal challenges associated with their design and implementation [198]. Based on existing conceptual models that describe factors of trust in automation [199], this section will overview best practices and recommendations that can render affective speech technologies trustworthy.

1) Preserving Privacy and Security: Affect in speech coexists with various types of personal identifiable information (PII), such as one's identity, age, perceived gender, and health status [200]. Thus, data and models pertaining to affective speech technologies might lead to unintended leaking of PII without the user's knowledge or consent [201]. The increasing public awareness of digital exposure [202] has motivated researchers to pursue research on privacy preservation in speech processing to ensure that the processing and analysis of one's voice occur such that undesired or unintended information

cannot be derived by the user, the system, or another third party [201], [203]. One potential approach to enhance PII preservation is via voice anonymization that typically encodes the voice characteristics of the speaker, the information that is spoken (i.e., linguistic content), and the co-occurring paralinguistic information that can be related to affect (i.e., intonation, rhythm, and vocal stress). These methods aim to alter the speaker's identity while preserving the linguistic and paralinguistic information [204], [205]. However, disentangling speaker information from segmental and suprasegmental information is not always straightforward since the latter is heavily affected by the idiosyncratic characteristics of one's voice. Initial studies have employed voice conversion to transform components of speech such as the spectral envelope [206]. Other approaches involve end-to-end systems that learn transformations of the speech waveform or spectrogram so that they minimize the automatic speech recognition (ASR) loss and maximize the speaker classification loss, implemented via adversarial learning approaches [207], [208], [209]. Recent research also focuses on replacing the identity of the original speaker with that of another speaker [204], [210], [211].

A limited number of recent studies have evaluated the ability of privacy-preserving anonymization techniques to preserve affect information. For example, Nortel et al. [212] replaced the speaker identity in one's voice with another speaker and found that this can preserve emotion information, yielding 15% degradation in emotion classification accuracy. Dias et al. [213] investigated hashing and homomorphic speech transformations that resulted in a 2%-3% degradation in classifying between neutral and angry speech. Feng and Narayanan [214] examined data transformation methods that allow inference of only desired affective attributes while suppressing others, as well as adaptive noise injection methods to suppress other PII attributes, such as gender from affect labels [215]. Tsouvalas et al. [216] studied a federated learning approach that was found to effectively remove speaker-dependent information while preserving the emotion recognition accuracy. Adversarial learning has been examined both in terms of learning privacy-enhanced emotion representations [217] and generating anonymized speech signals while preserving their emotional information [218]. Feng and Narayanan [137] proposed an effective set up of federated learning for SER, including differential privacy schemes to protect against attribute inference attacks [219].

It is becoming increasingly clear that privacy preservation should be a guiding principle when designing and deploying affective speech technologies. Yet, the interplay between PII and affect in speech prevents many of the existing and frequently used methods, such as adversarial learning, to succeed in the task of affect-preserving privacy mitigation. In addition, elements of privacy preservation should be considered along the well-known "personalization-privacy" paradox, which suggests that personalized services require users to continuously provide a wide variety of sensitive personal data, which increases their concerns about loss of privacy [220]. While prior work indicates a number of factors affecting users' tendency to share their data, such as past privacy violations and users' need for control [221], additional research at the intersection of engineering, engineering technology management, and human factors is required to better understand how to balance the personalization-privacy relationship and how these two factors influence users' acceptance of speech-based affect recognition technologies.

2) Ensuring Inclusiveness and Equity: Evidence by recent studies indicates that machine learning algorithms can learn sociodemographic attributes from the data, even if those are not explicitly encoded in the input of the algorithm [222], [223], [224]. This can systematically discriminate against certain subgroups of individuals, particularly ones belonging to sensitive populations (e.g., racially minoritized groups, older adults, and non-English speakers) for which enough data are not always available. In the case of speech, information about language/dialectal background and biological sex is inherently embedded in those signals; therefore, speech-based emotion recognition can naturally learn such sensitive information and use it in biased ways when taking decisions. Sensitive groups might depict distinct distribution from their counterparts due to biological reasons. For example, prior work has found significant main effects for biological sex and race in certain vocal tract dimensions [225], which can impact the corresponding acoustic measures. One's talking style might also be different depending on their socioeconomic status (SES) and social identity [226] that are associated with sociodemographic attributes. Sensitive groups may further employ different technological equipment (e.g., microphones) when recording speech in ambulatory settings, potentially due to differences in SES. This can result in the diverse quality of recordings that might bias the learned speech representations. Finally, the majority of work in SER typically focuses on high-resource languages, such as English, Mandarin, French, and German. Thus, inclusiveness should be also considered in the context of low-resource languages that do not typically receive much attention by the community and are rarely included in benchmarks, either via training ML models from scratch or by applying transfer learning and metalearning approaches [227], [228].

Given this evidence, the interest within the speech community in developing equitable speech-processing technologies has increased over the years. Here, the term "equitable" does *not only* refer to nondiscriminatory treatment but to a treatment that ensures that all individuals receive the same opportunity to reach a specific objective. Prior work suggests the presence of sociodemographic bias in ASR [229], [230], and voice biometrics and speaker recognition [224], [231]. Particularly, for the

case of speech-based emotion recognition, recent work has explored the effect of age and gender on the automated recognition of emotional valence and activation [222], [232]. Sagha et al. [222] found that it is easier to recognize valence in younger speakers, followed by female speakers. Gorrostieta et al. [232] found that it is more difficult to recognize emotion activation in females compared to male speakers and investigated various methods (i.e., adversarial learning, fairness-aware optimization) for debiasing the data. Gender bias has been examined in terms of acoustic, linguistic, and visual measures for emotion recognition [223], with text and audio being the least biased unimodal models, and video being the most biased, especially for happiness and frustration. Booth et al. [233] obtained similar results when using multimodal measures for the task of job hireability estimation based on video data; acoustic measures were found to be the least biased between female and male speakers, followed by linguistic and visual measures.

Despite the recent interest in equitable speech technologies, existing work on these technologies has been mostly developed in engineering labs and in isolation from the people and communities that they serve. It is important to adopt a holistic view and iterative design when building affective speech technologies. First and foremost, it is necessary to involve stakeholders in all stages of the model-building process so that their feedback is incorporated along the way. It is also important to rely on diverse research and development teams who can bring different perspectives based on their own background and experiences. Finally, it is necessary to monitor for potential biases throughout all stages of the technology-building process (i.e., data collection, analysis, model design, training, evaluation, and deployment) and rectify those when needed. Additional recommendations on creating inclusive AI technologies for human-centered applications can be found in [234] and [235].

3) Promoting Explainability for Calibrating User Trust: The large-scale deployment of speech emotion systems is bounded by the stakeholder's ability to explain the corresponding technology, especially when it comes to sensitive human-centered tasks, such as the ones pertaining to mental health, team science, and education. While much research has been done on designing explainable AI (XAI) and evaluating XAI with respect to user needs, this work is mostly focused on general machine learning that has been evaluated predominantly in computer vision and natural language processing [236], [237], [238]. In addition, the majority of work on explainability involves low-level technical explanations that might not be meaningful or relatable to the stakeholders. In tandem with XAI developments, it is important to consider recent findings from the field of human–computer interaction (HCI) that advocates for new XAI paradigms, such as ones that leverage contrastive explanations with counterfactual examples and contextual information [106], and incorporate value-sensitive design and participatory design [239]. Explainable emotion speech technologies should be developed in an iterative manner by clearly defining the system goals and design requirements (e.g., what to explain) and evaluating the system outcomes with the target users (e.g., clinician and patient) in terms of system usability and performance, human-XAI performance, overall user understanding, and appropriate user trust calibration [240]. In particular, for SER, a reasonable approach would be to integrate knowledge-driven information that reflects anticipated changes in speech characteristics resulting from the emotional content [241].

4) Considering the Ethical Implications: As affective speech technologies advance, a critical issue to address is the ethical and moral challenges associated with those, a domain of inquiry that is still in its infancy. Grounded in work from normative and applied ethics, research and development should be concerned about giving these technologies ethical principles and enabling them to function in a responsible manner for identifying and potentially resolving ethical dilemmas that are met as part of their decision-making process [242]. Since affective speech technologies are dealing with sensitive human-centered challenges, researchers designing and implementing those systems should be at the forefront of making sure that their design maximizes benefits, minimizes harm to the users, and abides with the morals of society. As the field progresses, it is important for the affective computing community to work with researchers from ethics and philosophy for identifying the ethical principles and guidelines that would help mitigate these issues. It is also necessary to raise public awareness of legitimate concerns, engage in conversations with the stakeholders (e.g., users and industry), and support policy makers toward a societal and legal framework in which the end-users can keep pace with the current developments and are not inadvertently negatively affected by those. The research community should further be proactive in getting involved in regulatory and legislatory efforts, such as the European Union's (EU) AI Act (AIA) [243] and the "Blueprint for an AI Bill of Rights" by the U.S. White House [244], in order to assist in creating an inclusive process for self-certification and government oversight of AI algorithms, define transparency requirements, and identify "acceptable" and "unacceptable" qualities and performance thresholds of the AI systems.

V. CONCLUSION

This article summarized the history and current state-of-the-art of technologies for recognizing emotions expressed in speech (i.e., SER). We presented findings from speech production research on how emotion influences the neuromuscular control of the vocal organs and, thus, speech articulation patterns and discussed how these findings have influenced the design of acoustic parameters that quantify the effect of emotion on speech phonation and articulation. Following that, we outlined the efforts of the

research community on curating and annotating emotion speech corpora and the challenges associated with this procedure, especially when it comes to capturing expressions and perceptions of emotions in real life. We also summarized modeling efforts that rely on knowledge-driven feature design and more recent representation learning approaches via deep learning. Noting advances in the field, moving from laboratory measurement to representing emotion expressed/perceived in natural real-world "in-the-wild" settings, we highlighted promising applications in health and education domains underscoring the importance of context-specific emotional speech data to be collected and modeled. Finally, we outlined challenges and engineering approaches related to responsibly and ethically integrating these technologies in real-life applications through a variety of means for ensuring trustworthiness, such as preserving a user's privacy and mitigating PII leaks, promoting their explainability to stakeholders, and enabling these technologies to work in an equitable manner for all people. ■

Acknowledgment

The authors would like to thank Dr. Daniel Bone for the valuable contributions and feedback on this manuscript.

REFERENCES

[1] P. Ekman, "Expression and the nature of emotion," *Approaches Emotion*, vol. 3, no. 19, p. 344, 1984.

[2] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. Face Gesture*, Mar. 2011, pp. 827–834.

[3] P. M. Desmet, M. H. Vastenburg, and N. Romero, "Mood measurement with pick-a-mood: Review of current methods and design of a pictorial self-report scale," *J. Design Res.*, vol. 14, no. 3, pp. 241–279, 2016.

[4] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, and E. P. Scilingo, "Automatic analysis of speech f0 contour for the characterization of mood changes in bipolar patients," *Biomed. Signal Process. Control*, vol. 17, pp. 29–37, Mar. 2015.

[5] J. Gideon, E. M. Provost, and M. McInnis, "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2359–2363.

[6] M. Hamilton, "A rating scale for depression," *J. Neurol., Neurosurg., Psychiatry*, vol. 23, no. 1, p. 56, 1960.

[7] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, "Young mania rating scale," in *Handbook of Psychiatric Measures*. Washington, DC, USA: American Psychiatric Association, 2000, pp. 540–542.

[8] S. Blanton, "The voice and the emotions," *Quart. J. Speech*, vol. 1, no. 2, pp. 154–172, 1915.

[9] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *J. Acoust. Soc. Amer.*, vol. 52, no. 4B, pp. 1238–1250, Oct. 1972.

[10] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Social Psychol.*, vol. 70, no. 3, pp. 614–636, 1996.

[11] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, Apr. 2003.

[12] G. Fairbanks and W. Pronovost, "Vocal pitch during simulated emotion," *Science*, vol. 88, no. 2286, pp. 382–383, Oct. 1938.

[13] D. Erickson, O. Fujimura, and B. Pardo, "Articulatory correlates of prosodic control: Emotion and emphasis," *Lang. Speech*, vol. 41, nos. 3–4, pp. 399–417, Jul. 1998.

[14] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *Proc. Interspeech*, Sep. 2005, pp. 1–4.

[15] D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya, "Exploratory study of some acoustic and articulatory characteristics of sad speech," *Phonetica*, vol. 63, no. 1, pp. 1–25, Mar. 2006.

[16] J. Kim, A. Toutios, S. Lee, and S. S. Narayanan, "A kinematic study of critical and non-critical articulators in emotional speech production," *J. Acoust. Soc. Amer.*, vol. 137, no. 3, pp. 1411–1429, Mar. 2015.

[17] J. Kim, A. Toutios, S. Lee, and S. S. Narayanan, "Vocal tract shaping of emotional speech," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101100.

[18] P. Kuppens, J. Stouten, and B. Mesquita, "Individual differences in emotion components and dynamics: Introduction to the special issue," *Cognition Emotion*, vol. 23, no. 7, pp. 1249–1258, Nov. 2009.

[19] R. J. Davidson, "Affective style and affective disorders: Perspectives from affective neuroscience," *Cognition Emotion*, vol. 12, no. 3, pp. 307–330, May 1998.

[20] K. R. Scherer, "Nonlinguistic vocal indicators of emotion and psychopathology," in *Emotions in Personality and Psychopathology*. Cham, Switzerland: Springer, 1979, pp. 493–529.

[21] A. Paeschke and W. F. Sendlmeier, "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements," in *Proc. ISCA Tutorial Res. Workshop (ITRW) Speech Emotion*, 2000, pp. 1–6.

[22] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Commun.*, vol. 16, no. 4, pp. 369–390, Jun. 1995.

[23] E. Rank and H. Pirker, "Generating emotional speech with a concatenative synthesizer," in *Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP)*, Nov. 1998, vol. 98, no. 3, pp. 671–674.

[24] R. W. Frick, "Communicating emotion: The role of prosodic features," *Psychol. Bull.*, vol. 97, no. 3, pp. 412–429, May 1985.

[25] K. Koike, H. Suzuki, and H. Saito, "Prosodic parameters in emotional speech," in *Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP)*, Nov. 1998, pp. 1–4.

[26] M. Schröder, "Emotional speech synthesis: A review," in *Proc. 7th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 2001, pp. 1–6.

[27] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current Directions Psychol. Sci.*, vol. 8, no. 2, pp. 53–57, Apr. 1999.

[28] L. Leinonen, T. Hiltunen, I. Linnankoski, and M.-L. Laakso, "Expression of emotional–motivational connotations with a one-word utterance," *J. Acoust. Soc. Amer.*, vol. 102, no. 3, pp. 1853–1863, Sep. 1997.

[29] C. Busso and S. S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Proc. ISCA Interspeech*, 2008, pp. 257–260.

[30] C. Dara, L. Monetta, and M. D. Pell, "Vocal emotion processing in Parkinson's disease: Reduced sensitivity to negative emotions," *Brain Res.*, vol. 1188, pp. 100–111, Jan. 2008.

[31] H. M. Gray and L. Tickle-Degnen, "A meta-analysis of performance on emotion recognition tasks in Parkinson's disease," *Neuropsychology*, vol. 24, no. 2, pp. 176–191, 2010.

[32] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *J. Cross-Cultural Psychol.*, vol. 32, no. 1, pp. 76–92, Jan. 2001.

[33] K. R. Scherer, E. Clark-Polner, and M. Mortillaro, "In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion," *Int. J. Psychol.*, vol. 46, no. 6, pp. 401–435, Dec. 2011.

[34] S. Hareli, K. Kafetsios, and U. Hess, "A cross-cultural study on emotion expression and the learning of social norms," *Frontiers Psychol.*, vol. 6, p. 1501, Oct. 2015.

[35] N. Lim, "Cultural differences in emotion: Differences in emotional arousal level between the east and the west," *Integrative Med. Res.*, vol. 5, no. 2, pp. 105–109, Jun. 2016.

[36] A. Peräkylä and M.-L. Sorjonen, *Emotion in Interaction*. Oxford, U.K.: Oxford Univ. Press, 2012.

[37] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Atlanta, GA, USA, 2006, p. 8.

[38] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[39] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition of emotions within speech," in *Proc. ISCA Interspeech*, 2007, pp. 2249–2252.

[40] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, Apr. 2013, pp. 1–8.

[41] M. Valstar et al., "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*. New York, NY, USA: ACM, Nov. 2014, pp. 3–10.

[42] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

[43] C. Eroglu Erdem, C. Turan, and Z. Aydin, "BAUM-2: A multilingual audio-visual affective face database," *Multimedia Tools Appl.*, vol. 74, no. 18, pp. 7429–7459, Sep. 2015.

[44] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan. 2017.

[45] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Lang. Resour. Eval.*, vol. 50, no. 3, pp. 497–521, Sep. 2016.

[46] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, 2016, pp. 1–9,

doi: 10.1109/WACV.2016.7477679.
[47] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct. 2019.
[48] H. Chou, W. Lin, L. Chang, C. Li, H. Ma, and C. Lee, "NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 292–298.
[49] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, "MEC 2017: Multimodal emotion recognition challenge," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact. (ACII Asia)*, May 2018, pp. 1–5.
[50] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "RAMAS: Russian multimodal corpus of dyadic interaction for affective computing," in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds. Cham, Switzerland: Springer, 2018, pp. 501–510.
[51] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 527–536.
[52] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*. Melbourne, VIC, Australia: Association for Computational Linguistics, 2018, pp. 2236–2246.
[53] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "SUST Bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for Bangla," *PLoS ONE*, vol. 16, no. 4, pp. 1–27, Apr. 2021, doi: 10.1371/journal.pone.0250173.
[54] J. Kossaifi et al., "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021.
[55] W. Yu et al., "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.
[56] H. Pérez-Espinosa, J. Martínez-Miranda, I. Espinosa-Curiel, J. Rodríguez-Jacobo, L. Villaseñor-Pineda, and H. Avila-George, "IESC-child: An interactive emotional children's speech corpus," *Comput. Speech Lang.*, vol. 59, pp. 55–74, Jan. 2020.
[57] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "DEMoS: An Italian emotional speech corpus: Elicitation methods, machine learning, and perception," *Lang. Resour. Eval.*, vol. 54, no. 2, pp. 341–383, Jun. 2020.
[58] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Proc. Interspeech*, Oct. 2020, pp. 1823–1827.
[59] J. Zhao et al., "M3ED: Multi-modal multi-scene multi-label emotional dialogue database," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 5699–5710.
[60] M. Jaiswal, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost, "MuSE: A multimodal dataset of stressed emotion," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2020, pp. 1499–1510.
[61] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.
[62] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digit. Signal Process.*, vol. 110, Mar. 2021, Art. no. 102951.
[63] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. M. Provost, "The PRIORI emotion dataset: Linking mood to emotion detected in-the-wild," in *Proc. ISCA Interspeech*, 2018, pp. 1903–1907.
[64] S. Leem, D. Fulford, J. Onnela, D. Gard, and C. Busso, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6447–6451.
[65] A. Wilf and E. M. Provost, "Towards noise robust speech emotion recognition using dynamic layer customization," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2021, pp. 1–8.
[66] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," in *Proc. Interspeech*, Sep. 2019, pp. 3148–3152.
[67] C. Oates, A. Triantafyllopoulos, I. Steiner, and B. W. Schuller, "Robust speech emotion recognition under different encoding conditions," in *Proc. Interspeech*, Sep. 2019, pp. 3935–3939.
[68] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.
[69] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.
[70] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Int. Congr. Acoust.*, vol. 19, no. 1, 2013, Art. no. 035081.
[71] (2022). *Pyaudio*. [Online]. Accessed: Mar. 14, 2023. Available: https://people.csail.mit.edu/hubert/pyaudio/
[72] (2022). *Pedalboard*. Accessed: Mar. 14, 2023. [Online]. Available: https://github.com/spotify/pedalboard
[73] M. Jaiswal and E. Mower Provost, "Best practices for noise-based augmentation to improve the performance of emotion recognition 'in the wild,'" 2021, *arXiv:2104.08806*.
[74] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychol. Bull.*, vol. 128, no. 2, pp. 203–235, 2002.
[75] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.
[76] M. Jaiswal et al., "Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7415–7419.
[77] K. P. Truong, D. A. van Leeuwen, and F. M. G. de Jong, "Speech-based recognition of self-reported and observed emotion in a dimensional space," *Speech Commun.*, vol. 54, no. 9, pp. 1049–1063, Nov. 2012.
[78] B. Zhang and E. M. Provost, "Automatic recognition of self-reported and perceived emotions," in *Multimodal Behavior Analysis in the Wild*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 443–470.
[79] J. Fahrenberg, M. Myrtek, K. Pawlik, and M. Perrez, "Ambulatory assessment–monitoring behavior in daily life settings," *Eur. J. Psychol. Assessment*, vol. 23, no. 4, pp. 206–213, Jan. 2007.
[80] T. J. Trull and U. Ebner-Priemer, "Ambulatory assessment," *Annu. Rev. Clin. Psychol.*, vol. 9, pp. 151–176, Nov. 2013.
[81] G. Kirouac and U. Hess, "Group membership and the decoding of nonverbal behavior," in *The Social Context of Nonverbal Behavior*, P. Philippot, R. Feldman, and E. Coats, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1999, pp. 182–210.
[82] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan. 2012.
[83] L. Devillers, R. Cowie, J.-C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie, "Real life emotions in French and English TV video clips: An integrated annotation protocol combining continuous and discrete approaches," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2006, pp. 1–6.
[84] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 76–89, Jan. 2018.
[85] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, Apr. 2015.
[86] G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "Deep canonical time warping for simultaneous alignment and representation learning of sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1128–1138, May 2018.
[87] S. Khorram, M. G. McInnis, and E. M. Provost, "Jointly aligning and predicting continuous emotion annotations," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1069–1083, Oct. 2021.
[88] A. Ramakrishna, R. Gupta, and S. Narayanan, "Joint multi-dimensional model for global and time-series annotations," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 473–484, Jan. 2022.
[89] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 890–897.
[90] H. Chou, W. Lin, C. Lee, and C. Busso, "Exploiting Annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7717–7721.
[91] B. Schuller et al., "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. INTERSPEECH*. Antwerp, Belgium: ISCA, 2007.
[92] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers Psychol.*, vol. 4, p. 292, Jan. 2013.
[93] C. M. Lee et al., "Emotion recognition based on phoneme classes," in *Proc. Interspeech*, Oct. 2004, pp. 1–6.
[94] D. Bone, C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: A rule-based framework withKnowledge-inspired vocal features," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 201–213, Apr. 2014.
[95] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.
[96] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1459–1462.
[97] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 960–964.
[98] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.
[99] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*,

Mar. 2016, pp. 5200–5204.
[100] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.
[101] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," 2019, *arXiv:1904.03833*.
[102] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 3465–3469.
[103] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 29, pp. 3451–3460, 2021.
[104] C. Lee, K. Sridhar, J. Li, W. Lin, B. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 22–38, Nov. 2021.
[105] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 992–1004, Apr. 2022.
[106] W. Zhang and B. Y. Lim, "Towards relatable explainable AI with the perceptual process," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2022, pp. 1–24.
[107] Y. Li, P. Bell, and C. Lai, "Fusing ASR outputs in joint training for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7362–7366.
[108] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7342–7346.
[109] T. Feng and S. Narayanan, "Semi-FedSER: Semi-supervised learning for speech emotion recognition on federated learning using multiview pseudo-labeling," in *Proc. Interspeech*, Sep. 2022, pp. 5050–5054.
[110] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, Aug. 2021, pp. 3400–3404.
[111] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Proc. Interspeech*, Aug. 2021, pp. 4508–4512.
[112] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[113] S. Latif, S. Khalifa, R. Rana, and R. Jurdak, "Federated learning for speech emotion recognition applications," in *Proc. Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, 2020, pp. 341–342.
[114] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7179–7183.
[115] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[116] G. Paraskevopoulos, E. Tzinis, N. Ellinas, T. Giannakopoulos, and A. Potamianos, "Unsupervised low-rank representations for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 939–943.
[117] J.-H. Tao, J. Huang, Y. Li, Z. Lian, and M.-Y. Niu, "Semi-supervised ladder networks for speech emotion recognition," *Int. J. Autom. Comput.*, vol. 16, no. 4, pp. 437–448, Aug. 2019.
[118] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. ISCA Interspeech*, 2018, pp. 3107–3111.
[119] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5099–5103.
[120] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 951–955.
[121] F. Tao and G. Liu, "Advanced LSTM: A study about better time dependency modeling in emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2906–2910.
[122] G. S. Malhi, A. Hamilton, G. Morris, Z. Mannie, P. Das, and T. Outhred, "The promise of digital mood tracking technologies: Are we heading on the right track?" *Evidence Based Mental Health*, vol. 20, no. 4, pp. 102–107, Nov. 2017.
[123] S. L. Bartels et al., "A narrative synthesis systematic review of digital self-monitoring interventions for middle-aged and older adults," *Internet Interventions*, vol. 18, Dec. 2019, Art. no. 100283.
[124] A. Ortiz and P. Grof, "Electronic monitoring of self-reported mood: The return of the subjective?" *Int. J. Bipolar Disorders*, vol. 4, no. 1, pp. 1–8, Dec. 2016.
[125] National Institute of Mental Health (NIMH). (2022). *Depression*. Accessed: Mar. 6, 2023. [Online]. Available: https://www.nimh.nih.gov/health/topics/depression
[126] E. Kraepelin, *Manic-Depressive Insanity and Paranoia*. Livingston, AL, USA: E.&S. Livingstone, 1921.
[127] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal assessment of depression from behavioral signals," in *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition*, vol. 2. San Rafael, CA, USA: Morgan & Claypool, 2018, pp. 375–417.
[128] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, Jul. 2015.
[129] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 3123–3128.
[130] F. Ringeval et al., "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, Oct. 2017, pp. 3–9.
[131] F. Ringeval et al., "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, Oct. 2019, pp. 3–12.
[132] E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96–107, Jan. 2008.
[133] L.-S.-A. Low, M. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in Adolescents' speech during family interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2011.
[134] K. E. B. Ooi, M. Lech, and N. B. Allen, "Multichannel weighted speech classification system for prediction of major depression in adolescents," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 2, pp. 497–506, Feb. 2013.
[135] B. Stasak, J. Epps, N. Cummins, and R. Goecke, "An investigation of emotional speech in depression classification," in *Proc. Interspeech*, Sep. 2016, pp. 485–489.
[136] K. Matton, M. G. McInnis, and E. M. Provost, "Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder," in *Proc. Interspeech*, Sep. 2019, pp. 1438–1442.
[137] A. C. Arevian et al., "Clinical state tracking in serious mental illness through computational analysis of speech," *PLoS ONE*, vol. 15, no. 1, Jan. 2020, Art. no. e0225695.
[138] R. Gupta et al., "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, Nov. 2014, pp. 33–40.
[139] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2016, pp. 35–42.
[140] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *J. Psychiatric Res.*, vol. 27, no. 3, pp. 309–319, Jul. 1993.
[141] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, and A. Othmani, "AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis," *Mach. Learn. Appl.*, vol. 2, Dec. 2020, Art. no. 100005.
[142] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7159–7163.
[143] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "SpeechFormer: A hierarchical efficient framework incorporating the characteristics of speech," in *Proc. Interspeech*, Sep. 2022, pp. 346–350.
[144] American Psychological Association. (2022). *APA Dictionary of Psychology—Autism Spectrum Disorder*. Accessed: Mar. 6, 2023. [Online]. Available: https://dictionary.apa.org/autism-spectrum-disorder?_ga=2.212234576.511431535.1678133050-1213189474.1664475545
[145] L. D. Shriberg, R. Paul, J. L. McSweeny, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and prosody characteristics of adolescents and adults with high-functioning autism and asperger syndrome," *J. Speech, Lang., Hearing Res.*, vol. 44, no. 5, pp. 1097–1115, Oct. 2001.
[146] C. Trevarthen and S. Daniel, "Disorganized rhythm and synchrony: Early signs of autism and rett syndrome," *Brain Develop.*, vol. 27, pp. S25–S34, Nov. 2005.
[147] K. L. Cooper and T. L. Hanstock, "Confusion between depression and autism in a high functioning child," *Clin. Case Stud.*, vol. 8, no. 1, pp. 59–71, Feb. 2009.
[148] D. Bone et al., "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist," in *Proc. Interspeech*, Sep. 2012, pp. 1043–1046.
[149] D. Bone et al., "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *J. Speech, Lang., Hearing Res.*, vol. 57, no. 4, pp. 1162–1177, Aug. 2014.
[150] B. Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, Aug. 2013, pp. 148–152.
[151] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller, "Speech-based diagnosis of autism spectrum condition by generative adversarial network representations," in *Proc. Int. Conf. Digit. Health*, Jul. 2017, pp. 53–57.
[152] J. H. Lee, G. W. Lee, G. Bong, H. J. Yoo, and H. K. Kim, "Deep-Learning-Based detection of infants with autism spectrum disorder using auto-encoder feature representation," *Sensors*, vol. 20, no. 23, p. 6762, Nov. 2020.
[153] F. B. Pokorny et al., "Earlier identification of

children with autism spectrum disorder: An automatic vocalisation-based approach," in *Proc. Interspeech*, Aug. 2017, pp. 309–313.

[154] N. J. Chambers, A. M. Wetherby, S. T. Stronach, N. Njongwe, S. Kauchali, and R. R. Grinker, "Early detection of autism spectrum disorder in young isiZulu-speaking children in south Africa," *Autism*, vol. 21, no. 5, pp. 518–526, Jul. 2017.

[155] A. Khozaei, H. Moradi, R. Hosseini, H. Pouretemad, and B. Eskandari, "Early screening of autism spectrum disorder using cry features," *PLoS ONE*, vol. 15, no. 12, Dec. 2020, Art. no. e0241690.

[156] T. Sorensen, E. Zane, T. Feng, S. Narayanan, and R. Grossman, "Cross-modal coordination of face-directed gaze and emotional speech production in school-aged children and adolescents with ASD," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019.

[157] S. B. Gaigg, "The interplay between emotion and cognition in autism spectrum disorder: Implications for developmental theory," *Frontiers Integrative Neurosci.*, vol. 6, p. 113, Jan. 2012.

[158] J. M. Garcia-Garcia, V. M. R. Penichet, M. D. Lozano, and A. Fernando, "Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions," *Universal Access Inf. Soc.*, vol. 21, no. 4, pp. 809–825, Nov. 2022.

[159] National Institute of Agining. (2022). *Parkinson's Disease: Causes, Symptoms, and Treatments*. Accessed: Mar. 7, 2023. [Online]. Available: https://www.nia.nih.gov/health/parkinsons-disease

[160] J. R. Orozco-Arroyave, *Analysis of Speech of People With Parkinson's Disease*, vol. 41. Berlin, Germany: Logos Verlag, 2016.

[161] E. Schalling, K. Johansson, and L. Hartelius, "Speech and communication changes reported by people with Parkinson's disease," *Folia Phoniatrica et Logopaedica*, vol. 69, no. 3, pp. 131–141, 2017.

[162] J. R. Orozco-Arroyave et al., "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6490–6494.

[163] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22199–22208, 2017.

[164] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[165] T. Bocklet, E. Nöth, G. Stemmer, H. Ruzickova, and J. Rusz, "Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2011, pp. 478–483.

[166] M. S. Barnish, S. M. C. Horton, Z. R. Butterfint, A. B. Clark, R. A. Atkinson, and K. H. O. Deane, "Speech and communication in Parkinson's disease: A cross-sectional exploratory study in the UK," *BMJ Open*, vol. 7, no. 5, May 2017, Art. no. e014642.

[167] L. X. Blonder, R. E. Gur, and R. C. Gur, "The effects of right and left hemiparkinsonism on prosody," *Brain Lang.*, vol. 36, no. 2, pp. 193–207, Feb. 1989.

[168] C. Schröder, Z. T. Nikolova, and R. Dengler, "Changes of emotional prosody in Parkinson's disease," *J. Neurological Sci.*, vol. 289, nos. 1–2, pp. 32–35, Feb. 2010.

[169] M. Pell, H. Cheang, and C. Leonard, "The impact of Parkinson's disease on vocal-prosodic communication from the perspective of listeners," *Brain Lang.*, vol. 97, no. 2, pp. 123–134, May 2006.

[170] S. Zhao, F. Rudzicz, L. G. Carvalho, C. Marquez-Chin, and S. Livingstone, "Automatic detection of expressed emotion in Parkinson's disease," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4813–4817.

[171] Y. Kim, T. Soyata, and R. F. Behnagh, "Towards emotionally aware AI smart classroom: Current issues and directions for engineering and education," *IEEE Access*, vol. 6, pp. 5308–5331, 2018.

[172] M. Yadav, M. N. Sakib, E. H. Nirjhar, K. Feng, A. H. Behzadan, and T. Chaspari, "Exploring individual differences of public speaking anxiety in real-life and virtual presentations," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1168–1182, Jul. 2022.

[173] R. Southwell et al., "Challenges and feasibility of automatic speech recognition for modeling Student collaborative discourse in classrooms," *Thinking*, vol. 27, pp. 29–42, 2022.

[174] M. E. Dale, A. J. Godley, S. A. Capello, P. J. Donnelly, S. K. D'Mello, and S. P. Kelly, "Toward the automated analysis of teacher talk in secondary ELA classrooms," *Teaching Teacher Edu.*, vol. 110, Feb. 2022, Art. no. 103584.

[175] L. Eloy et al., "Modeling team-level multimodal dynamics during multiparty collaboration," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 244–258.

[176] C. Sun et al., "The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment," *Comput. Hum. Behav.*, vol. 128, Mar. 2022, Art. no. 107120.

[177] P. Paromita, A. Khader, S. Begerowski, S. T. Bell, and T. Chaspari, "Linguistic and vocal markers of microbehaviors between team members during analog space exploration missions," *IEEE Pervasive Comput.*, early access, Jan. 31, 2023, doi: 10.1109/MPRV.2022.3232780.

[178] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annu. Rev. Psychol.*, vol. 66, pp. 799–823, Jan. 2015.

[179] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, Nov. 2018.

[180] N. van Berkel, M. B. Skov, and J. Kjeldskov, "Human-AI interaction: Intermittent, continuous, and proactive," *Interactions*, vol. 28, no. 6, pp. 67–71, Nov. 2021.

[181] J. Jiang, A. J. Karran, C. K. Coursaris, P.-M. Léger, and J. Beringer, "A situation awareness perspective on human-AI interaction: Tensions and opportunities," *Int. J. Hum.–Comput. Interact.*, vol. 39, pp. 1–18, Jun. 2022.

[182] A. A. Tutul, E. H. Nirjhar, and T. Chaspari, "Investigating trust in human-machine learning collaboration: A pilot study on estimating public anxiety from speech," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 288–296.

[183] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.

[184] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1055–1068, Oct. 2021.

[185] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," in *Proc. Interspeech*, Aug. 2017, pp. 1098–1102.

[186] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 85–99, Jan. 2019.

[187] K. Feng and T. Chaspari, "Few-shot learning in emotion recognition of spontaneous speech using a Siamese neural network with adaptive sample pair formation," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1627–1633, Jun. 2021.

[188] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "EmoNet: A transfer learning framework for multi-corpus speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, pp. 1472-1487, 2021.

[189] L. K. Barr, J. H. Kahn, and W. J. Schneider, "Individual differences in emotion expression: Hierarchical structure and relations with psychological distress," *J. Social Clin. Psychol.*, vol. 27, no. 10, pp. 1045–1077, Dec. 2008.

[190] L. A. King and R. A. Emmons, "Conflict over emotional expression: Psychological and physical correlates," *J. Personality Social Psychol.*, vol. 58, no. 5, pp. 864–877, 1990.

[191] W. Ng and E. Diener, "Personality differences in emotions: Does emotion regulation play a role?" *J. Individual Differences*, vol. 30, no. 2, pp. 100–106, Jan. 2009.

[192] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 5117–5120.

[193] J.-L. Li and C.-C. Lee, "Attentive to individual: A multimodal emotion recognition network with personalized attention profile," in *Proc. ISCA Interspeech*, 2019, pp. 211–215.

[194] N. Vryzas, L. Vrysis, R. Kotsakis, and C. Dimoulas, "A Web crowdsourcing framework for transfer learning and personalized speech emotion recognition," *Mach. Learn. with Appl.*, vol. 6, Dec. 2021, Art. no. 100132.

[195] Z. N. Karam et al., "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4858–4862.

[196] S. Khorram, J. Gideon, M. McInnis, and E. M. Provost, "Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge," in *Proc. Interspeech*, Sep. 2016, pp. 1215–1219.

[197] S. Yan, H. Hosseinmardi, H. Kao, S. Narayanan, K. Lerman, and E. Ferrara, "Estimating individualized daily self-reported affect with wearable sensors," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2019, pp. 1–9.

[198] S. Day, C. Seninger, J. Fan, K. Pundi, A. Perino, and M. Turakhia, "Digital health consumer adoption report 2019," Rock Health, San Francisco, CA, USA, Tech. Rep., 2019.

[199] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 46, no. 1, pp. 50–80, 2004.

[200] S. Rukavina, S. Gruss, H. Hoffmann, J.-W. Tan, S. Walter, and H. C. Traue, "Affective computing and the impact of gender and age," *PLoS ONE*, vol. 11, no. 3, Mar. 2016, Art. no. e0150584.

[201] T. Feng, R. Hebbar, N. Mehlman, X. Shi, A. Kommineni, and A. S. Narayanan, "A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness," 2022, *arXiv:2212.09006*.

[202] L. Rainie et al., "Anonymity, privacy, and security online," *Pew Res. Center*, vol. 5, Sep. 2023.

[203] M. A. Pathak, B. Raj, S. D. Rane, and P. Smaragdis, "Privacy-preserving speech processing: Cryptographic and string-matching frameworks show promise," *IEEE Signal Process. Mag.*, vol. 30, no. 2, pp. 62–74, Mar. 2013.

[204] F. Fang et al., "Speaker anonymization using X-vector and neural waveform models," in *Proc. ISCA Interspeech*, 2019, pp. 155–160.

[205] N. Tomashenko et al., "The VoicePrivacy 2020 challenge: Results and findings," *Comput. Speech Lang.*, vol. 74, Jul. 2022, Art. no. 101362.

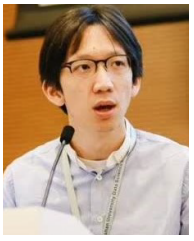
[206] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," in *Proc. ISCA Interspeech*, 2020, pp. 1099–1103.

[207] F. Bahmaninezhad, C. Zhang, and J. Hansen, "Convolutional neural network based speaker de-identification," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, Jun. 2018, pp. 255–260.

[208] P. Champion, D. Jouvet, and A. Larcher, "Speaker information modification in the VoicePrivacy 2020 toolchain," in *Proc. ISCA Interspeech*, 2020,

pp. 1693–1697.
[209] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, "Speaker anonymization for personal information protection using voice conversion techniques," *IEEE Access*, vol. 8, pp. 198637–198645, 2020.
[210] B. M. L. Srivastava et al., "Design choices for X-vector based speaker anonymization," in *Proc. Interspeech*, Oct. 2020, pp. 1713–1717.
[211] B. M. L. Srivastava et al., "Privacy and utility of X-vector based speaker anonymization," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 30, pp. 2383–2395, 2022.
[212] H. Nourtel, P. Champion, D. Jouvet, A. Larcher, and M. Tahon, "Evaluation of speaker anonymization on emotional speech," in *Proc. ISCA Symp. Secur. Privacy Speech Commun.*, Nov. 2021, pp. 1–9.
[213] M. Dias, A. Abad, and I. Trancoso, "Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2057–2061.
[214] T. Feng and S. Narayanan, "Privacy and utility preserving data transformation for speech emotion recognition," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2021, pp. 1–7.
[215] T. Feng, H. Hashemi, M. Annavaram, and S. S. Narayanan, "Enhancing privacy through domain adaptive noise injection for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7702–7706.
[216] V. Tsouvalas, T. Ozcelebi, and N. Meratnia, "Privacy-preserving speech emotion recognition through semi-supervised federated learning," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops Affiliated Events (PerCom Workshops)*, Mar. 2022, pp. 359–364.
[217] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 7985–7993.
[218] V. Ravuri, R. Gutierrez-Osuna, and T. Chaspari, "Preserving mental health information in speech anonymization," in *Proc. 10th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Oct. 2022, pp. 1–8.
[219] T. Feng, R. Peri, and S. Narayanan, "User-level differential privacy against attribute inference attack of speech emotion recognition on federated learning," in *Proc. Interspeech*, Sep. 2022, pp. 5055–5059.
[220] X. Guo, X. Zhang, and Y. Sun, "The privacy–personalization paradox in mHealth services acceptance of different age groups," *Electron. Commerce Res. Appl.*, vol. 16, pp. 55–65, Mar. 2016.
[221] G. N. Punj, "Understanding individuals' intentions to limit online personal information disclosures to protect their privacy: Implications for organizations and public policy," *Inf. Technol. Manage.*, vol. 20, no. 3, pp. 139–151, Sep. 2019.
[222] H. Sagha, J. Deng, and B. Schuller, "The effect of personality trait, age, and gender on the performance of automatic speech valence recognition," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 86–91.
[223] M. Schmitz, R. Ahmed, and J. Cao, "Bias and fairness on multimodal emotion detection algorithms," 2022, *arXiv:2205.08383*.
[224] R. Peri, K. Somandepalli, and S. Narayanan, "A study of bias mitigation strategies for speaker recognition," *Comput. Speech Lang.*, vol. 79, Apr. 2023, Art. no. 101481.
[225] S. A. Xue and J. G. Hao, "Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry," *J. Voice*, vol. 20, no. 3, pp. 391–400, Sep. 2006.
[226] S. E. Gaither, A. M. Cohen-Goldberg, C. L. Gidney, K. B. Maddox, C. L. Gidney, and C. L. Gidney, "Sounding black or white: Priming identity and biracial speech," *Frontiers Psychol.*, vol. 6, p. 457, Apr. 2015.
[227] F. Haider, S. Pollak, P. Albert, and S. Luz, "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods," *Comput. Speech Lang.*, vol. 65, Jan. 2021, Art. no. 101119.
[228] S. Chopra, P. Mathur, R. Sawhney, and R. R. Shah, "Meta-learning for low-resource speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6259–6263.
[229] A. Koenecke et al., "Racial disparities in automated speech recognition," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020.
[230] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," 2021, *arXiv:2103.15122*.
[231] X. Chen, Z. Li, S. Setlur, and W. Xu, "Exploring racial and gender disparities in voice biometrics," *Sci. Rep.*, vol. 12, no. 1, pp. 1–12, Mar. 2022.
[232] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2823–2827.
[233] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 268–277.
[234] M. K. Lee and K. Rich, "Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–14.
[235] A. C. Timmons et al., "A call to action on assessing and mitigating bias in artificial intelligence applications for mental health," *Perspect. Psychol. Sci.*, vol. 9, Dec. 2022, Art. no. 17456916221134490.
[236] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[237] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–42, Sep. 2019.
[238] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 8, pp. 2674–2693, Aug. 2019.
[239] U. Ehsan and M. O. Riedl, "Human-centered explainable AI: Towards a reflective sociotechnical approach," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2020, pp. 449–466.
[240] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, Dec. 2021.
[241] K. Feng, "Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Sep. 2022, pp. 1–7.
[242] S. L. Anderson, "Machine metaethics," *Mach. ethics*, vol. 21, p. 27, Jun. 2011.
[243] T. Burri and F. von Bothmer. (2022). *The New EU Legislation on Artificial Intelligence: A Primer*. Accessed: Mar. 6, 2023. [Online]. Available: http://dx.doi.org/10.2139/ssrn.3831424
[244] *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*, White House Office Sci. Technol. Policy (OSTP), Washington, DC, USA, 2022.

ABOUT THE AUTHORS



Chi-Chun Lee (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2007 and 2012, respectively.

He is currently a Professor with the Department of Electrical Engineering, National Tsing Hua University (NTHU), Hsinchu, Taiwan. He holds joint appointments in Communication Engineering, Semiconductor Research, BioAI, and Precision Medicine Ph.D. Programs. He has published over 160 peer-reviewed publications and is a coauthor on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, and IEEE Engineering in Medicine and Biology Conference (EMBC) 2020. His research interests are in speech and language, affective computing, health analytics, and behavioral signal processing.

Dr. Lee is a TPC Member of the APSIPA Image, Video, and Multimedia (IVM) and Machine Learning and Data Analytics (MLDA) Committees. He was a recipient of the Foundation of Outstanding Scholar's Young Innovator Award in 2020, the CIEE Outstanding Young Electrical Engineer Award in 2020, the IICM K. T. Li Young Researcher Award in 2020, the NTHU Industry Collaboration Excellence Award in 2021, and the MOST Futuretek Breakthrough Awards in 2018 and 2019. He has served as the Area Chair of Interspeech 2016, 2018, and 2019, the Senior Program Committee Chair of Affective Computing & Intelligent Interaction (ACII) 2017 and 2019, the Publicity Chair of ACM International Conference on Multimodal Interaction (ICMI) 2018, the Late Breaking Result Chair of ACM ICMI 2023, and the Sponsorship and Special Session Chair of International Symposium on Chinese Spoken Language Processing (ISCSLP) 2018 and 2020. He is the General Chair of IEEE Automatic Speech Recognition and Understanding (ASRU) 2023. He was an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA from 2019 to 2020. He has been an Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING since 2020, the *Journal of Computer Speech and Language* since 2021, and the *APSIPA Transactions on Signal and Information Processing*.

Theodora Chaspari (Member, IEEE) received the B.S. degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 2010, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2012 and 2017, respectively.

She is currently an Assistant Professor of computer science and engineering with Texas A&M University, College Station, TX, USA. Her research interests lie in human-centered machine learning and affective computing and are supported by the National Science Foundation (NSF), national indsitutes of health (NIH), national aeronautics and space administration (NASA), air force office of scientific research (AFOSR).

Dr. Chaspari was a recipient of the TEES Dean of Engineering Excellence Award in 2022, the NSF CAREER Award in 2021, the TAMU Montague Teaching Award in 2021, and the USC Women in Science and Engineering Merit Fellowship in 2015. Papers coauthored with her students have been nominated and won awards in IEEE Affective Computing & Intelligent Interaction (ACII) 2019, ACM BuildSys 2019, ASCE International Conference on Computing in Civil Engineering (i3CE) 2019, and IEEE Body Sensor Networks (BSN) 2018. She is also serving as an Editor for the *Computer Speech and Language* (Elsevier) and a Guest Editor for the IEEE Transactions on Affective Computing.

Emily Mower Provost (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2010.

She is currently an Associate Professor of computer science and engineering with the University of Michigan, Ann Arbor, MI, USA. Her research interests are in human-centered speech and video processing, multimodal interface design, and speech-based assistive technology. The goals of her research are motivated by the complexities of the perception and expression of human behavior.

Dr. Provost received best paper awards or finalist nominations for Interspeech 2008, ACM Multimedia 2014, International Conference on Multimodal Interaction (ICMI) 2016, and IEEE Transactions on Affective Computing. She was a Toyota Faculty Scholar in 2020. She has been awarded the National Science Foundation CAREER Award in 2017, the Oscar Stern Award for Depression Research in 2015, and the National Science Foundation Graduate Research Fellowship from 2004 to 2007. Among other organizational duties, she has been the Program Chair of Affective Computing & Intelligent Interaction (ACII) in 2017 and 2021 and ICMI in 2016 and 2018. She is an Associate Editor of IEEE Transactions on Affective Computing and the IEEE Open Journal of Signal Processing. She has also served as an Associate Editor for *Computer Speech and Language* and *ACM Transactions on Multimedia*.

Shrikanth S. Narayanan (Fellow, IEEE) was with AT&T Bell Labs/AT&T Research, NY. He is currently a University Professor and the Niki & C. L. Max Nikias Chair in Engineering with the University of Southern California (USC), Los Angeles, CA, USA, where he holds appointments as a Professor of electrical and computer engineering, computer science, linguistics, psychology, neuroscience, otolaryngology, and pediatrics, the Research Director of the Information Science Institute, and the Director of the Ming Hsieh Institute. At USC, he leads the Signal Analysis and Interpretation Laboratory (SAIL), with a research focus on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biosignal processing, machine intelligence, and their applications with direct societal relevance. His research and inventions have led to technology commercialization including through startups that he cofounded: Behavioral Signals Technologies, CA, focused on the telecommunication services and AI-based conversational assistance industry and Lyssn focused on mental health care delivery, treatment, and quality assurance.

Dr. Narayanan is a Guggenheim Fellow, a member of the European Academy of Sciences and Arts, and a Fellow of the National Academy of Inventors, the Acoustical Society of America, International Speech and Communication Association (ISCA), the American Association for the Advancement of Science, the Association for Psychological Science, the Association for the Advancement of Affective Computing, and the American Institute for Medical and Biological Engineering. He is a recipient of several honors, including the 2023 ISCA Medal for Scientific Achievement; the 2015 Engineers Council's Distinguished Educator Award; the Mellon Award for Mentoring Excellence; the 2005 and 2009 Best Journal Paper Awards from the IEEE Signal Processing Society and served as its Distinguished Lecturer from 2010 to 2011; the 2018 ISCA CSL Best Journal Paper Award and served as an ISCA Distinguished Lecturer from 2015 to 2016 and a Willard R. Zemlin Memorial Lecturer for ASHA in 2017; the Ten Year Technical Impact Award in 2014; and the Sustained Accomplishment Award in 2020 from ACM International Conference on Multimodal Interaction (ICMI). He has served as the Inaugural VP-Education of the IEEE Signal Processing Society from 2020 to 2022. He is an Editor of the *Computer Speech and Language* journal. He has served as the Editor-in-Chief of the IEEE Journal of Selected Topics in Signal Processing.