Anusandhanvallari
Vol 2026, No.1
October 2026
ISSN 2229-3388

Emotion Beyond Words: Deep Learning Approaches to Detect Sentiment and Tone in Multilingual Speech

Dr Aliyaparveen A Mulla[1], Dr Mallanna Biradar[2], Daisy Kalita[3], Ms. Nanda Makhija[4], Akansh Garg[5]

[1]Associate Professor, Department of English, Government First Grade College Managuli, Vijayapura, Managuli, Karnataka

[2]Associate Professor, Department of English, MGVC Arts Commerce and Science College Muddebihal, Vijayapur, Muddebihal, Karnataka

[3]Assistant Professor, Computer Science, University of Science & Technology Meghalaya, Ri Bhoi, Khanapara, Meghalaya,

[4]Assistant Professor, Department of English, JSPM University, Pune, Maharashtra

[5]Director Array Research Pvt Ltd

Abstract: Speech carries sentiment and emotion in its words and also in pitch, loudness, rhythm, and voice quality, and these cues are shaped by language and culture. Systems that detect sentiment and tone from speech must therefore work across languages, speakers, and recording conditions, yet most were built and tested on a few acted English or German corpora. This paper reviews the empirical evidence on deep-learning approaches to multilingual speech emotion and sentiment recognition. It combines emotion theory and cross-cultural perception studies with work on acoustic features, end-to-end networks, self-supervised multilingual encoders, benchmark corpora, multimodal fusion, fairness, and regulation. Four findings emerge. First, self-supervised encoders have changed the field. Multilingual encoders are pretrained on up to 436,000 hours of speech in 128 languages, whereas the standard benchmark corpus, IEMOCAP, holds about 12 hours from ten actors, and fine-tuned transformers have narrowed the long-standing gap in valence prediction. Second, people recognize vocal emotion across cultures at above-chance rates but with an in-group advantage, and machine systems show a parallel pattern, with accuracy falling under cross-corpus and cross-lingual testing. Third, evaluation practice limits what can be concluded: on a 300-utterance test set, 70 percent accuracy carries a 95 percent interval of about ±5 points, wider than many reported gains. Fourth, the risks are concrete. Commercial speech recognizers had an average word error rate of 0.35 for Black speakers against 0.19 for white speakers, and the EU AI Act prohibits emotion inference in workplaces and educational institutions, except for medical or safety reasons. Two comparative tables map methods to evidence and challenges to mitigations. The paper then proposes a calibrated multilingual pipeline with an evaluation protocol. The pipeline is a conceptual synthesis of the reviewed literature and has not been tested experimentally.

Keywords: Speech Emotion Recognition, Multilingual Speech, Sentiment Analysis, Self-Supervised Learning, Cross-Lingual Transfer, Paralinguistics, Algorithmic Fairness.

I. Introduction

Three related terms need separating. Sentiment is usually a polarity (positive, negative, or neutral), emotion is a category or a position on continuous dimensions, and tone is how something is said, which can signal attitude independently of the words. Schuller's review of two decades of speech emotion recognition describes a field that moved from small acted corpora and hand-designed features toward larger data and learned representations [1]. The acoustic basis is long established: Banse and Scherer showed that vocal portrayals of different emotions have

distinguishable acoustic profiles [2]. Russell's circumplex model describes emotion by valence and arousal, which is the basis for dimensional labeling [3].

Deep learning changed the modeling. Trigeorgis and coauthors trained a convolutional recurrent network directly on raw waveforms and showed that it could replace hand-crafted features for dimensional emotion prediction [4]. Self-supervised models then learned general speech representations from unlabeled audio [5]. Wagner and coauthors fine-tuned such transformers on a naturalistic corpus and reported that they improved valence prediction, historically the weakest dimension for audio alone, while apparently learning linguistic information implicitly [6].

Multilingual scale is the next step. XLS-R was pretrained on 436,000 hours of speech in 128 languages [7], and Whisper was trained on 680,000 hours of labeled speech, of which about 117,000 hours cover 96 languages other than English [8]. IEMOCAP, the benchmark most used for emotion recognition, contains about 12 hours of data [32], so the pretraining corpus of XLS-R is more than 36,000 times larger (436,000 divided by 12, a derived comparison). Pretraining supplies acoustic and phonetic knowledge across languages, but emotion labels remain scarce and are unevenly distributed across languages.

The central difficulty is generalization. Schuller and coauthors showed that emotion recognizers trained on one corpus lose substantial accuracy when tested on another, because corpora differ in speakers, recording conditions, emotion elicitation, and labeling [9]. Neumann and Vu extended the question to languages, studying English and French [10]. A multilingual deployment, such as a contact center serving callers in many Indian and international languages, faces all of these shifts at once.

This paper reviews how deep-learning systems address sentiment and tone in multilingual speech, how they are evaluated, and where the evidence is weak.

Research Objectives

- To review theory and cross-cultural evidence on how emotion is expressed and perceived in speech.
- To compare feature-based, end-to-end, self-supervised, and multilingual approaches to speech emotion and sentiment recognition.
- To assess the corpora and evaluation practices on which reported results rest.
- To evaluate evidence on cross-corpus and cross-lingual transfer and on multimodal fusion with text.
- To identify fairness, privacy, and regulatory constraints, and to propose a pipeline and research priorities.

II. Literature Review

2.1 Emotion Theory and Cross-Cultural Perception

Ekman's basic-emotions view holds that a small set of emotions are universal [11]. Barrett and coauthors argue, in the context of facial movements, that inferring emotion from expression is much less reliable than commonly assumed because expression varies with context and culture [12]. For speech systems, this means that labels record what annotators perceived or what actors were told to portray, not an observed inner state.

Perception studies show both universality and cultural tuning. Elfenbein and Ambady's meta-analysis, which covered 97 studies, found that emotions are recognized above chance across cultures but more accurately when judge and expresser share a cultural group [13]. Scherer, Banse, and Wallbott had listeners in nine countries judge vocal expressions by German actors and found above-chance recognition everywhere, with accuracy lower for listeners whose languages were more distant from German [14]. Pell and coauthors compared English, German, Hindi, and Arabic and found that emotions were recognized well above chance in all four, with differences by

emotion and listener language [15]. Laukka and coauthors studied expression and recognition across five nations, including India, and found largely shared acoustic cues alongside an in-group advantage [16]. Cowen and coauthors found that listeners in the United States and India could distinguish twelve categories of emotion in speech prosody, supporting both shared categories and cultural variation [17].

Language adds a specific complication. In tone languages, such as Mandarin, pitch distinguishes word meanings [18], so fundamental frequency (F0), the main prosodic cue to emotion, serves lexical and affective functions at once. A model that learned "high pitch means excitement" from a non-tonal language can misread a tonal one.

On the acoustic side, the Geneva Minimalistic Acoustic Parameter Set standardized a small feature set, 62 parameters in the basic version and 88 in the extended version, covering pitch, loudness, spectral balance, and voice-quality measures, so that studies would be comparable [19]. These features remain a useful interpretable baseline and diagnostic.

2.2 From Handcrafted Features to End-to-End Learning

Earlier reviews describe a pipeline of acoustic features fed to classifiers such as support vector machines or hidden Markov models, applied to small acted corpora [20], [21]. Mirsamadi and coauthors introduced a recurrent network with local attention that learned to weight the emotionally salient frames of an utterance and improved over a feature-based baseline on IEMOCAP [22]. Trigeorgis et al. removed hand-crafted features altogether [4]. Latif and coauthors' survey of deep representation learning concluded that learned representations generally outperform hand-designed ones, while noting that data scarcity, cross-corpus robustness, and label quality remain open problems [23]. Because labeled data are small, augmentation is standard; SpecAugment, which masks bands of time and frequency in the spectrogram, was developed for speech recognition and is widely adopted for emotion tasks [24].

2.3 Self-Supervised and Multilingual Encoders

Self-supervised models pretrain on unlabeled speech by predicting masked or quantized representations, then transfer to downstream tasks. wav2vec 2.0 learns contextual representations at about 50 frames per second (a 20 ms stride), so a 5-second utterance becomes about 250 frames [5]. HuBERT predicts clustered hidden units for masked frames [25]. WavLM extends this with denoising and was pretrained on 94,000 hours of audio [26]. The SUPERB benchmark compares such encoders across tasks with a frozen encoder and light task heads, and includes emotion recognition on IEMOCAP, where the large WavLM model reached roughly 70 percent accuracy in the emotion task [26], [27].

Layer choice matters. Pepino and coauthors used wav2vec 2.0 embeddings with learned weights over hidden layers for emotion recognition and reported gains over spectral baselines on IEMOCAP, indicating that intermediate layers carry much of the affective information [28]. Wagner et al. fine-tuned transformer encoders on MSP-Podcast and reported gains on arousal, dominance, and valence, with the largest change in valence [6]. Emotion-specific pretraining is also available: emotion2vec reports improvements over general-purpose self-supervised models on IEMOCAP and on corpora in other languages [29].

Multilingual coverage is addressed by pretraining on many languages, as with XLS-R [7], and by multilingual benchmarks. EmoBox standardizes data splits and evaluation across dozens of emotion corpora in 14 languages, so that models can be compared under common conditions [30]. For Indian languages, the IndicSUPERB benchmark provides about 1,684 hours of labeled speech across 12 languages, though for recognition and speaker tasks and not emotion [31]. To the extent that public emotion-labeled corpora in Indian languages are fewer than the available speech data, emotion systems for these languages will depend on transfer. Whisper's encoder is a possible feature extractor [8], but the reviewed benchmark evidence on its use for emotion is limited.

2.4 Corpora and Evaluation

Corpus properties determine what results mean. IEMOCAP contains about 12 hours of dyadic interactions from ten actors in five sessions, partly scripted and partly improvised [32]. RAVDESS contains 7,356 recordings from 24 professional actors, of which 1,440 are speech audio files [33]. EMO-DB has 535 utterances from ten actors in German across seven emotions [34]. CREMA-D has 7,442 clips from 91 actors [35]. MSP-Podcast is drawn from podcast recordings and labeled for both categories and dimensions, which makes it more naturalistic [36]. Most of these are acted, in a small number of languages, with few speakers, so speaker identity can be confounded with emotion unless splits are speaker-independent.

Metrics need care. The INTERSPEECH 2009 challenge promoted unweighted average recall, UAR = (1/K) Σ recall_k over K classes, because class frequencies are unequal [37]. As an illustration, a common four-class IEMOCAP configuration contains 5,531 utterances: 1,708 neutral, 1,636 happy or excited, 1,103 angry, and 1,084 sad. A model that always predicts neutral achieves weighted accuracy of 1,708 ÷ 5,531 = 30.9 percent and UAR of 25 percent. Weighted accuracy overstates ability on imbalanced data. For dimensional prediction, the concordance correlation coefficient, $\rho c = 2\rho\sigma_x\sigma_y / (\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2)$, penalizes both low correlation and bias in mean and scale [38].

Sample size limits resolution. The standard error of an accuracy p on n independent test items is $\sqrt{(p(1 - p)/n)}$. At p = 0.70 and n = 1,000, the 95 percent interval is about ±2.8 points; at n = 300, about ±5.2 points. Because test utterances from the same speaker are correlated, true intervals are wider. Differences of one or two points between systems on small corpora are therefore not strong evidence of improvement.

2.5 Cross-Corpus and Cross-Lingual Transfer

Schuller et al. showed that performance falls substantially when training and test corpora differ, and examined strategies such as normalization and training on pooled corpora to reduce the loss [9]. Neumann and Vu studied English and French and reported reduced accuracy when a model trained in one language was tested in the other, with joint multilingual training a feasible way to recover performance [10]. Latif and coauthors examined Urdu against Western-language corpora and reported that transfer to Urdu was weak in some language pairs and improved when some target-language data were included [39]. These machine results mirror the human in-group advantage [13], [14]: models do worst on languages and cultures furthest from their training data.

2.6 Multimodal Fusion and Text Sentiment

Words carry sentiment that audio does not. Zadeh and coauthors' Tensor Fusion Network combined language, acoustic, and visual features and outperformed unimodal baselines on a multimodal sentiment benchmark [40]. CMU-MOSEI contributed about 23,500 English sentences from over 1,000 YouTube speakers, labeled for sentiment and emotion [41]. MELD provided about 13,000 utterances from multi-party dialogues in a television series, labeled with seven emotions and sentiment [42]. Tsai and coauthors' multimodal transformer used cross-modal attention so that audio and text streams need not be word-aligned [43]. Both major multimodal corpora are English, so evidence for multilingual fusion is thin.

Text sentiment benefits from multilingual language models. XLM-R was pretrained on text in 100 languages and improved cross-lingual transfer [44]. XLM-T adapted this to Twitter and released a sentiment benchmark in eight languages, including Hindi [45]. A cascade system transcribes speech and applies a text model, but errors compound. As an illustration, if recognition errors affect each sentiment-bearing word independently with probability 0.2, an utterance with three such words keeps all of them correct with probability $0.8^3 = 51.2$ percent. Cascades also discard prosody, so sarcasm or flat delivery of positive words is missed.

2.7 Fairness, Privacy, and Regulation

Speech technologies perform unevenly across groups. Koenecke and coauthors tested five commercial speech recognition systems and found an average word error rate of 0.35 for Black speakers against 0.19 for white speakers [46]. A cascade that depends on these transcripts inherits the disparity. Gorrostieta and coauthors showed that emotion recognizers can carry gender bias and used adversarial training to reduce it, with a cost in overall performance [47].

Regulation is tightening. The EU Artificial Intelligence Act prohibits AI systems that infer the emotions of people in workplaces and educational institutions, except for medical or safety purposes, and treats other emotion-recognition systems as high-risk, with prohibitions applying from February 2025 [48]. India's Digital Personal Data Protection Act, 2023 governs the processing of digital personal data on a consent basis and does not address emotion inference specifically [49]. Health uses are sensitive: Cummins and coauthors reviewed speech analysis for depression and suicide-risk assessment and found promising acoustic markers, but concluded that robust clinical use requires larger and more diverse data [50].

2.8 Research Gap

Studies typically report accuracy on one or a few corpora in one language, with speaker-dependent splits or small test sets, and few evaluate across languages, report uncertainty, or test subgroup performance. This review compares the evidence on a common footing and proposes a pipeline and protocol.

III. Methodology

This study is a qualitative, comparative review with a conceptual synthesis step. Studies were included if they (i) addressed emotion, sentiment, or tone recognition from speech, or the speech representations and corpora used for it, or related perception, fairness, or regulatory questions; (ii) reported empirical results on named corpora or controlled perception studies, or were official regulatory documents; and (iii) appeared in a peer-reviewed journal or proceedings, or were issued by a legislature. Studies were compared along four axes: modeling approach; corpus and languages; evaluation protocol (speaker independence, cross-corpus testing, metric); and reported finding. Reported figures are those of the original authors, are not re-run, and are not comparable across studies because corpora, label sets, and splits differ. Illustrative calculations (the majority-class baseline, confidence intervals, cascade error, and the risk–coverage example) use stated or assumed inputs and are labeled as such. No new experiments are reported. Because most corpora are English, German, or Mandarin, claims for Indian and other under-resourced languages are extrapolations.

IV. Results and Discussion

4.1 Approaches, Data, and Reported Findings

Table 1 maps the principal approaches onto data and reported findings.

Table 1. Approaches to Speech Emotion and Sentiment Recognition: Data and Reported Findings

Approach	Representative Work	Data	Reported Finding	Ref.
Handcrafted acoustic features	GeMAPS / eGeMAPS	Multiple corpora	Standard 62- and 88-parameter sets for comparability	[19]
End-to-end convolutional-recurrent	Trigeorgis et al.	Raw waveform, dimensional labels	Replaced hand-crafted features	[4]

Recurrent network with attention	Mirsamadi et al.	IEMOCAP	Improved over feature-based baseline	[22]
Self-supervised encoders	wav2vec 2.0, HuBERT, WavLM	SUPERB, IEMOCAP	WavLM Large roughly 70% accuracy	[26], [27]
Layer-weighted embeddings	Pepino et al.	IEMOCAP	Gains over spectral baselines; intermediate layers informative	[28]
Fine-tuned transformers	Wagner et al.	MSP-Podcast	Improved arousal, dominance, valence; valence gap narrowed	[6]
Emotion-specific pretraining	emotion2vec	IEMOCAP and other languages	Improved over general self-supervised models	[29]
Multilingual pretraining	XLS-R	436,000 hours, 128 languages	Basis for cross-lingual transfer	[7]
Multilingual benchmark	EmoBox	Dozens of corpora, 14 languages	Standardized splits for comparison	[30]
Multimodal fusion	TFN, multimodal transformer	CMU-MOSEI and related	Fusion outperformed unimodal baselines	[40], [43]
Multilingual text sentiment	XLM-R, XLM-T	Text, 100 languages; Twitter, 8 languages	Cross-lingual text baselines	[44], [45]
Cross-lingual speech emotion	Neumann and Vu; Latif et al.	English–French; Urdu vs. Western	Accuracy falls across languages; target data helps	[10], [39]

4.2 Discussion: What the Evidence Supports

Table 1 shows that representation learning has delivered the clearest gains. Moving from hand-crafted features to learned representations, and then to self-supervised encoders pretrained on far more audio than any emotion corpus contains, improved results on every benchmark reported [4], [26], [28]. The Wagner results are notable because valence, the positive–negative dimension, is the hardest to read from acoustics alone, and fine-tuned transformers appeared to compensate by learning linguistic content implicitly [6]. That is a strength for benchmarks where speakers say meaningful sentences, and a risk elsewhere: a model that leans on words may fail in a language it has seen little of, or when words and tone conflict.

Multilingual encoders widen coverage without solving transfer. The pretraining of XLS-R covers 128 languages, but emotion labels exist for far fewer, and the cross-lingual studies show losses when the target language differs from the training set [7], [10], [39]. This is consistent with the human in-group advantage [13], [14]. A reasonable expectation, not yet established by direct evidence for most Indian languages, is that fine-tuning on even a modest amount of target-language labeled data will matter more than the choice of encoder.

Multimodal evidence is stronger for English than for other languages. Fusion helps where text and acoustics carry complementary information [40], [43], but no reviewed study tests it on multilingual corpora of the size of CMU-MOSEI or MELD [41], [42].

Evaluation is the weakest link. The majority-class example (30.9 percent weighted accuracy against 25 percent UAR) and the confidence-interval arithmetic (±5.2 points at 300 utterances) show that headline numbers on small, acted, speaker-overlapping corpora cannot support fine distinctions between systems [32], [37]. The SUPERB

figure of roughly 70 percent for IEMOCAP should be read as a four-class result on acted, English data, not as expected accuracy in the field.

4.3 Challenges, Evidence, and Mitigations

Table 2 sets the principal challenges against evidence and mitigations.

Table 2. Challenges in Multilingual Speech Emotion Recognition: Evidence and Mitigation

Challenge	Evidence Reviewed	Mitigation	Ref.
Ambiguous, perceived labels	Expression varies with context and culture	Dimensional labels; multiple annotators; report annotator language	[3], [12]
Small, acted corpora	IEMOCAP about 12 hours, 10 actors; EMO-DB 535 utterances	Naturalistic corpora; large-scale pretraining	[32], [34], [36]
Cross-corpus shift	Accuracy falls when corpus changes	Pooled training; normalization; cross-corpus testing	[9]
Cross-lingual shift	Losses when language changes; in-group advantage in people	Target-language fine-tuning; multilingual benchmarks	[10], [30], [39]
Tone languages	Pitch is lexical as well as affective	Speaker-normalized F0; language-aware evaluation	[18]
Cascade errors	Illustrative 51.2% chance three cue words survive 20% error	Audio-text fusion; use ASR confidence	[8], [43]
Unequal ASR accuracy	Word error rate 0.35 vs. 0.19	Subgroup evaluation; balanced data	[46]
Demographic bias in emotion models	Gender bias reduced by adversarial training	Debiasing; subgroup metrics	[47]
Small test sets and weak metrics	±5.2 points at n = 300; weighted accuracy misleading	UAR, CCC, confidence intervals, speaker-independent splits	[37], [38]
Regulation and privacy	Prohibition in workplace and education; consent rules	Restrict uses; consent; human oversight	[48], [49]

4.4 Discussion: Reliability, Fairness, and Permissible Use

Table 2 shows that most problems come from the mismatch between what is labeled and what is deployed. Labels reflect perceived emotion in a specific language and culture [12], [13], so a system that reports “angry” is reporting a prediction of what a certain group of annotators would call angry. For multilingual use, annotation by speakers of each target language is a requirement, not a refinement.

Fairness problems compound in cascades. A transcription error rate that is nearly twice as high for one group means that text-based sentiment is degraded more for that group [46]. Audio-only models avoid the transcript dependency but have their own demographic effects [47]. The practical response is to report performance by language, accent, gender, and age, not only overall.

The risk–coverage trade-off is a useful design tool. As an illustration with assumed numbers, suppose a system is 70 percent accurate at full coverage, so 30 of 100 utterances are misclassified. If it abstains on the 20 least confident utterances and accuracy on the remaining 80 rises to 77 percent, then 61.6 are correct and 18.4 are wrong, so errors fall from 30 to 18.4 per 100 inputs, with 20 sent to a human reviewer. Whether that is worthwhile depends on the cost of a wrong label compared with the cost of review. The example requires well-calibrated confidence scores, which are not guaranteed by the models reviewed.

Regulation limits permissible use independent of accuracy. Under the EU AI Act, inferring emotions in workplaces and educational institutions is prohibited except for medical or safety reasons, so a contact-center analytics product sold into the EU for monitoring employees would fall within the prohibition [48]. India's data-protection law places consent and purpose obligations on voice data without addressing emotion inference [49]. Health uses require clinical validation beyond what the benchmarks provide [50].

4.5 A Calibrated Multilingual Pipeline and Evaluation Protocol

The reviewed evidence suggests a design and evaluation protocol. It is a conceptual synthesis from the evidence above and has not been tested.

1. **Define the construct.** State whether the target is perceived emotion, sentiment polarity, or arousal and valence, and by whom it is perceived. Prefer dimensional labels where categories differ across languages [3], [12].
2. **Choose the encoder.** Use a multilingual self-supervised encoder, such as XLS-R, or an emotion-pretrained model, such as emotion2vec, with learned layer weights and fine-tuning on target-language data [7], [28], [29].
3. **Add a lexical channel.** Transcribe with a multilingual recognizer and feed the transcript to a multilingual text encoder, fusing with the acoustic stream, and weight the text channel by recognition confidence [8], [43], [44].
4. **Handle tone.** Normalize pitch per speaker and include language identity as a feature or stratum in evaluation, so that lexical tone is not mistaken for affect [18].
5. **Calibrate and abstain.** Calibrate output probabilities and route low-confidence cases to human review, using the risk–coverage trade-off to choose the threshold.
6. **Evaluate rigorously.** Use speaker-independent splits, cross-corpus and leave-one-language-out tests, UAR and CCC, and confidence intervals that account for speaker clustering, and report subgroup metrics by language, accent, gender, and age [9], [37], [38], [46].
7. **Govern use.** Document purpose and consent, exclude prohibited contexts, and require human oversight for consequential decisions [48], [49].

V. Conclusion

This paper reviewed deep-learning approaches to detecting sentiment and tone in multilingual speech. Three conclusions follow. First, learned representations have produced the main advances: self-supervised encoders pretrained on up to 436,000 hours of audio across 128 languages, and fine-tuned transformers that narrowed the valence gap, though benchmark results of roughly 70 percent come from acted English data [6], [7], [27]. Second, generalization across corpora and languages remains the central weakness, and it mirrors the in-group advantage found in human listeners [9], [10], [13]. Third, evaluation and governance lag modeling: small test sets give intervals of about ±5 points, transcription accuracy differs by group, and the EU AI Act restricts emotion inference

in workplaces and education [46], [48]. The practical recommendation is to treat outputs as calibrated estimates of perceived emotion, validate them in each target language and subgroup, and limit use to permitted contexts.

VI. Future Work

Future research should prioritize several areas. First, naturalistic emotion corpora annotated by native speakers are needed for Indian and other under-resourced languages, since existing corpora are concentrated in a few languages [30], [31]. Second, controlled comparisons of multilingual encoders under common splits should test whether target-language fine-tuning matters more than encoder choice [7], [29], [30]. Third, multilingual multimodal corpora of the size of CMU-MOSEI are needed to test audio–text fusion beyond English [41], [43]. Fourth, studies should evaluate calibration and abstention in emotion recognition, and measure the risk–coverage trade-off against human review costs. Fifth, fairness audits should report subgroup performance as a standard practice, extending work on ASR and gender bias [46], [47]. Finally, the pipeline in Section 4.5 should be implemented and tested on a multilingual corpus with speaker-independent and leave-one-language-out splits.

References

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, May 2018.

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

[3] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[4] 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)*, 2016, pp. 5200–5204.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 12449–12460.

[6] J. Wagner *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, Sep. 2023.

[7] A. Babu *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022, pp. 2278–2282.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 28492–28518.

[9] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul.–Dec. 2010.

[10] M. Neumann and N. T. Vu, "Cross-lingual and multilingual speech emotion recognition on English and French," in *Proc. IEEE ICASSP*, 2018, pp. 5769–5773.

[11] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.

[12] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Interest*, vol. 20, no. 1, pp. 1–68, Jul. 2019.

[13] 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.

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

[15] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *J. Phonetics*, vol. 37, no. 4, pp. 417–435, Oct. 2009.

[16] P. Laukka *et al.*, "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features," *J. Pers. Soc. Psychol.*, vol. 111, no. 5, pp. 686–705, 2016.

[17] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, "The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures," *Nature Hum. Behav.*, vol. 3, no. 4, pp. 369–382, Apr. 2019.

[18] M. Yip, *Tone*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[19] 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.–Jun. 2016.

[20] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[21] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020.

[22] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE ICASSP*, 2017, pp. 2227–2231.

[23] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1634–1654, Apr.–Jun. 2023.

[24] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[25] W.-N. Hsu *et al.*, "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.

[26] S. Chen *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[27] S. Yang *et al.*, "SUPERB: Speech processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[28] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[29] Z. Ma *et al.*, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Findings Assoc. Comput. Linguistics: ACL 2024*, 2024, pp. 15747–15760.

[30] Z. Ma *et al.*, "EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," in *Proc. Interspeech*, 2024, pp. 1580–1584.

[31] T. Javed *et al.*, "IndicSUPERB: A speech processing universal performance benchmark for Indian languages," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 11, 2023, pp. 12942–12950.

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

[33] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, Art. no. e0196391, May 2018.

[34] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

[35] 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.–Dec. 2014.

[36] 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.–Dec. 2019.

[37] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.

[38] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, Mar. 1989.

[39] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, 2018, pp. 88–93.

[40] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2017, pp. 1103–1114.

[41] A. 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 (ACL)*, 2018, pp. 2236–2246.

[42] 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 ACL*, 2019, pp. 527–536.

[43] Y.-H. H. Tsai *et al.*, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting ACL*, 2019, pp. 6558–6569.

[44] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting ACL*, 2020, pp. 8440–8451.

[45] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados, "XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond," in *Proc. 13th Lang. Resour. Eval. Conf. (LREC)*, 2022, pp. 258–266.

[46] A. Koenecke *et al.*, "Racial disparities in automated speech recognition," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020.

[47] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 2823–2827.

[48] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official J. Eur. Union*, Jul. 2024.

[49] Government of India, "The Digital Personal Data Protection Act, 2023," Act No. 22 of 2023, New Delhi, India, 2023.

[50] N. Cummins *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, Sep. 2015.