Enabling Internationalization of Affective Speech Technology using LLMs

Bo-Hao Su*, Shinji Watanabe* Chi-Chun Lee†
* Language Technologies Institute, Carnegie Mellon University, USA
E-mail: {bohaos, swatanab}@andrew.cmu.edu
† Electrical Engineering, National Tsing Hua University, Taiwan
E-mail: cclee@ee.nthu.edu.tw

***Abstract*—Affective speech technology aspires to equip machines with the ability to sense, interpret, and generate emotionally expressive speech, enabling empathetic assistants, social robots, and digital health companions. Large Audio/Speech Language Models (LALMs/SpeechLMs) now dominate this space: a single model can perform speech recognition, affect detection, and emotion-controlled synthesis, achieving impressive zero-shot generalization. However, we argue that LALMs are not yet internationalized: culturally grounded affect is misread when training data are skewed, leading to mis-recognition of affect, culturally inappropriate responses, and uneven user experiences.**

This paper surveys the current state of affective speech processing with LALMs, cataloging leading models, their sensing-to-synthesis capabilities, and the databases and metrics used for evaluation. We identify the key obstacle to responsible deployment: the heterogeneity of human vocal expression across cultures, which manifests as data scarcity, model bias, and evaluation blind spots. To address this gap, we propose a research agenda comprising: (i) systematic analysis of cultural variation in vocal affect, (ii) computational strategies for contextualizing LALMs toward culturally sensitive emotion processing, and (iii) benchmarks featuring balanced corpora and culture-aware metrics. By charting these directions, we aim to advance affective speech technology that is globally robust, socially responsible, and truly inclusive. The overall concept is depicted in Figure 1.

I. Introduction

Affective speech technology equips computational systems to perceive, interpret, and generate emotionally nuanced speech. This capability underpins applications such as empathetic voice assistants[1], [2], social robots[3], contact-center agents[4], mental-health support[5], and education[6], where alignment to human affect improves user engagement across signal processing, machine learning, psychology, and HCI.

The last few years have witnessed a paradigm shift toward LALMs/SpeechLMs as the de-facto state-of-the-art framework for speech intelligence, e.g., Qwen2Audio[7]. Leveraging billions of parameters and training corpora that blend speech with textual and multimodal resources, contemporary LALMs can jointly execute tasks that formerly required carefully tuned cascades: speech recognition, semantic understanding, affect recognition, and emotionally controlled speech synthesis. Their emergent zero-shot and few-shot generalization abilities dramatically reduce the engineering burden for new domains, while prompt-based conditioning enables fine-grained control over linguistic, paralinguistic, and affective attributes. Consequently, LALMs have become the principal technological foundation for both academic exploration and industrial deployment of affective speech solutions.

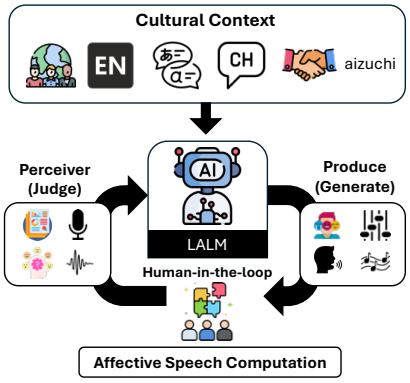

Fig. 1. Cultural-aware LALMs in affective speech computation

Notwithstanding these advances, the field now faces a critical internationalization challenge. Emotion is a culturally embedded construct: prosodic cues that denote politeness, respect, or enthusiasm can vary markedly across languages and societies[8], [9]. Current LALMs are predominantly trained on data skewed toward a small subset of high-resource languages and homogeneous recording conditions. This linguistic and cultural imbalance manifests as domain mismatch across speaker demographics, acoustic environments, and interactional styles, leading to systematically degraded performance or culturally inappropriate responses in under-represented communities. Such biases have practical consequences from eroding user satisfaction to perpetuating social inequities, and are particularly salient for the Asia-Pacific region targeted by APSIPA, where linguistic diversity is exceptionally high.

Against this backdrop, the present paper makes three primary contributions. First, it offers a comprehensive survey of recognition and generation techniques, tracing the evolution from single-language statistical methods to multimodal, multilingual LALMs, and highlighting open problems unique to affective processing. Second, it articulates a position that future progress hinges on rigorous internationalization, and hence on the development of balanced data resources, culturally aware evaluation metrics, and realistic task protocols that extend beyond conventional accuracy measures. Third, it proposes a roadmap for collaborative, human-in-the-loop methodologies, integrating cultural expertise with the adaptive strengths of

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

Stage	Company	LALMs	ASR (WER↓)	TTS (WER↓)	IEMOCAP (Acc.↑)	MSP-Podcast 1.7 (UAR↑)	MELD (Acc.↑)
Early	Meta	Spirit-LM[10]	21.9$^{\diamondsuit}$/29.2$^{\clubsuit}$	45.5$^{\diamondsuit}$/43.8$^{\clubsuit}$	-	-	-
	Google	AudioPaLM[11]	11.1(VoxPopuli[12])	-	-	-	-
Lately	Alibaba	Qwen2Audio	1.6$^{\diamondsuit}$/3.6$^{\clubsuit}$	-	59.2	-	55.3
	Aamazon	SpeechVerse[13]	2.1$^{\diamondsuit}$/4.4$^{\clubsuit}$	-	-	66.7	-
	ByteDance	SALMONN[14]	2.1$^{\diamondsuit}$/4.9$^{\clubsuit}$	-	69.0*	-	-
	NVIDIA	Audio Flamingo 3	1.57$^{\diamondsuit}$/3.13$^{\clubsuit}$	2.02 (SEED[15])	63.8	-	-
	Microsoft	Phi4-mm[16]	1.68$^{\diamondsuit}$/3.83$^{\clubsuit}$	-	-	-	-

TABLE I

RECENT LALMS PERFORMANCE REPORT ON ASR, TTS AND SER DOWNSTREAM TASKS ACCORDING TO ORIGINAL TECHNICAL PUBLIC REPORT, WHERE $\diamondsuit$, $\clubsuit$ REPRESENT FOR LIBRISPEECH TEST-CLEAN AND TEST-OTHER, AND $\star$ SIGN MEANS TEST ON SESSION 5 ONLY.

LALMs, to mitigate bias and enhance global applicability. By synthesizing current knowledge and setting forth concrete research directions, the paper aims to catalyze an inclusive next generation of affective speech technology that faithfully reflects the emotional diversity of its worldwide users.

II. BACKGROUND AND SCOPE

A. Current State of Affective Speech Computation

Conventional affective speech research has historically split downstream work into recognition and generation. In the recognition branch, models are trained to infer a speaker's emotional state, either in categorical terms, e.g., anger, joy, or along continuous dimensions such as valence and arousal across settings that range from monologues to dyadic and group dialogues [17]–[22]. The generation branch, by contrast, focuses on tasks such as emotional voice conversion and emotional text-to-speech, where systems synthesize speech that conveys a target affect [23]–[27]. As the field matured, researchers pushed beyond single-modality and single-language paradigms toward multimodal and multilingual emotion models[28]–[31]. This expansion exposed pronounced domain mismatch problems, differences in speaker demographics, recording conditions, languages, and input signal formats all degrade model robustness[32]–[35]. Studies that weave emotional awareness into human-computer interfaces underscore the importance of solving these issues: emotion-aware systems consistently improve user satisfaction and trust, highlighting affective speech technology's inherently interdisciplinary nature at the intersection of psychology, linguistics, computer science, and sociology[36]–[38].

Most recently, the emergence of LALMs trained on massive speech–text corpora has been a game-changer. A single LALM can unify what once required a cascade of specialized modules, simultaneously handling speech generation, comprehension, and emotion recognition. Yet, because their training data remain linguistically and culturally skewed, even the most capable LALMs inherit bias and struggle to generalize globally, reinforcing the need for culturally balanced corpora and evaluation frameworks if affective speech technology is to achieve true international reach.

B. Related Works

1) LALMs in Affective Computation: As noted above, contemporary LALMs extend pre-trained LLM backbones with lightweight acoustic adapters, enabling a single architecture to address a wide spectrum of downstream tasks, including automatic speech recognition (ASR), speech-to-text translation, speech-emotion recognition (SER), and spoken-language understanding. For instance, Qwen2Audio[7], trained on a blended corpus of speech, environmental sound, and music, attains **55.3%** accuracy on the MELD SER benchmark while achieving word-error rates (WER) of **1.6%** and **3.6%** on LibriSpeech *test-clean* and *test-other*, respectively. Likewise, NVIDIA's recent Audio Flamingo 3[39] reports **63.8%** accuracy on IEMOCAP and WERs of **1.57%/3.13%** on the same LibriSpeech splits. Furthermore, next-generation "duplex" LALMs accept a continuous user audio stream while emitting low-latency codec-based responses, employing channel-fusion layers that jointly model simultaneous user and agent speech[40], [41]. A broader comparison of representative models is provided in Table I with common tasks such as ASR, TTS and our focused SER. These results underscore the feasibility of *universal* LALMs that seamlessly transition from sensing (ASR, SER) to synthesis (emotion-controllable TTS) and even duplex listen–talk interaction. However, from Table I, it is obvious that excluding ASR, other downstream tasks lack unified comparisons, leading to unclear conclusion in SER according to these public technical reports.

To further enhance LALM affective perception while mitigating the cost of manual annotation, recent work explores data augmentation[42], automated labeling pipelines[43], [44], and zero-shot or weakly supervised adaptation strategies[45]. Collectively, these advances define the current frontier of affective speech processing with LALMs.

2) Measurement and Evaluation: Conventional SER evaluates emotion understanding by partitioning corpora such as IEMOCAP[48], MSP-Podcast[49], and MELD[50] with *leave-one-session-out* or *leave-one-speaker-out* schemes, and reporting class-balanced metrics, most commonly *unweighted average recall* (UAR), alongside weighted accuracy/recall, F1, precision, and recall. For dimensional labels, correlation and error measures, e.g., CCC, MAE, are standard. Contemporary LALMs typically follow these practices for *closed-form* classification with a fixed label set. Large-scale SER evaluation corpora are summarized in Table II.

Because LALMs are inherently *generative*, evaluation must also cover *open-form* outputs (free-text appraisals and rationales). Predictions are post-hoc mapped to a target ontology, via controlled verbalizers[51], lexicon/regex normalization, or rubric-based *LLM-as-adjudicator*, to mitigate label-space drift and taxonomy mismatch. Mapping rules should be pre-registered and paired with sensitivity analyses, since metrics

Corpus	Language	Setting	Modality	Hours	Emotions
MSP-Podcast v1.12	en	Podcast-in-the-wild	{Text, Speech}	∼324	Categorical/{Val, Aro, Dom}
BIIC-Podcast[46]	zh-tw	Podcast-in-the-wild	{Text, Speech}	∼147	Categorical/{Val, Aro, Dom}
LSSED[47]	en	Natural	{Speech}	∼206	Categorical
MELD	en	TV-show	{Text, Speech, Video}	∼13	Categorical
IEMOCAP	en	Dyadic (improvisation, script)	{Text, Speech, Video}	∼12	Categorical

TABLE II
LARGE-SCALE EMOTION CORPORA, WHERE VAL, ARO, AND DOM REPRESENTS VALENCE, AROUSAL AND DOMINANCE RESPECTIVELY.

can vary with normalization.

For affective *generation*, e.g., emotional TTS or conversion, evaluation combines objective and human judgements: naturalness (MOS/CMOS), affective correctness/strength (human raters or a held-out SER), and prosodic alignment to targets (F0 range, energy, speaking rate).

In summary, while traditional metrics remain essential for comparability, LALM-centric evaluation should explicitly separate closed vs. open form, standardize ontology mapping, and incorporate calibration, robustness, cross-corpus/lingual fairness, and human-centered appropriateness.

III. CHALLENGES FOR LALMS IN CULTURAL SENSITIVITY OF AFFECTIVE SPEECH

Humans' emotional perspectives and expressions are shaped by the languages and cultures in which they develop. Subtle socialization cues-display rules, politeness norms, backchannels, and conversational timing-accumulate over years, producing worldwide differences that form a diverse spectrum of emotions. Emotions are therefore not universal constructs; they are experienced, displayed, and interpreted differently across communities, making affective speech recognition and generation inherently culture-sensitive. As LALMs enter everyday applications, the imperative is *internationalization*: systems must be reliable across languages and cultures and produce contextually appropriate responses for diverse regions, an essential requirement for all-inclusive, responsible AI. In practice, this involves not only label prediction but also turn-timing, prosodic choices, and the justifications agents offer to users from different cultural backgrounds.

Dual roles of LALMs: production and perception. Beyond acting as task solvers, modern LALMs now serve (i) **producers** of content and data-synthesizing text, speech, and even images for augmentation or pretraining, and (ii) **perceivers/judges** scoring responses, adjudicating open-form outputs, and filtering/reranking candidates. *The model's "opinion" therefore shapes both sides of the pipeline*. If generative outputs are not culturally aligned, e.g., TTS that overuses high-arousal prosody in cultures that value low-arousal positivity, the resulting synthetic corpora entrench skewed norms. If an LLM acts as a judge with majority-culture priors, open-form appraisals and rationales from minority speakers can be systematically undervalued. Closed feedback loops, models training on, and being evaluated by, their own culturally biased outputs, amplify these effects.

Despite strong general capabilities, present LALMs may neglect culture-specific nuances that shape emotion. For example, [52] reports that East-Asian cultures are associated with greater use of suppression and avoidance than Western individuals. Similarly, Tsai [53], [54] shows that Americans tend to value high-arousal positive states (e.g., enthusiastic, excited) more, and low-arousal positive states (e.g., calm, relaxed, peaceful) less than East Asians. In everyday interaction, this surfaces as systematic errors: reserved prosody mistaken for indifference; honorific speech mapped to "joy"; tonal F0 patterns misread as arousal; Japanese *aizuchi* treated as positive affect rather than listener engagement; or sarcasm and indirect disagreement flattened into generic negativity. Such misalignments degrade user trust and the perceived appropriateness of system behaviour, and when the model is both producer and judge propagate into datasets and metrics.

These limitations arise naturally from the diversity and heterogeneity of spoken language as a function of cultural variation. Concretely, (i) *data imbalance* (long-tailed language/dialect coverage, sparse annotations for under-resourced cultures), (ii) *inconsistent ontologies* and annotation guidelines across corpora (category vs. valence–arousal taxonomies, display-rule instructions), (iii) *context loss* (missing metadata on speaker demographics, setting, relationship, and intent), and (iv) *acoustic/domain mismatch* (recording conditions, devices, interaction styles) create distribution shifts that current models fail to bridge. On the technology side, pretraining objectives are largely task-agnostic and text-centric; affect is weakly supervised; ASR biases propagate to downstream affect inference; decoding/control mechanisms lack explicit cultural priors; and duplex timing models insufficiently capture region-specific turn-taking norms. Critically, today's pipelines include too few *culturally aware human factors*: cultural experts are rarely embedded in data curation, rubric design, or adjudication, leaving production and perception stages to inherit majority-culture defaults.

Addressing these challenges requires a coordinated agenda centered on both roles of LALMs. For **production**, ensure culture-aligned generation via culturally conditioned prompts/control tokens, style-/prosody adapters, and expert-curated guardrails; release synthetic data with provenance and cultural metadata. For **perception/judging**, complement LLM-as-judge with culturally diverse expert panels, pre-registered rubrics, and cross-region calibration; avoid self-referential training/evaluation loops by triangulating with human annotations and independent models. More broadly, develop culturally grounded ontologies and balanced corpora; benchmarks that probe closed- and open-form SER as well as duplex generation; and evaluation beyond accuracy to fairness, calibration, robustness, appropriateness, and reasoning quality. Without such internationalization and without sustained human-in-the-loop engagement from cultural experts progress in affective

computing with LALMs will remain uneven and exclusionary.

IV. Future Directions

We take the position that cultural sensitivity must be engineered into both generation (production) and evaluation (perception). Current data-hungry LALMs lack systematic investigation and principled modeling of cultural factors during pretraining and adaptation. As a result, subtle but consequential nuances in prosody, pragmatics, and interactional norms are often sacrificed, limiting reliability across the world's diverse linguistic communities. We argue that a *computational and systematic* program for culturally sensitive adaptation is imperative: culturally adapted LALMs can deliver more appropriate responses, reduce misunderstanding in context-dependent scenarios, and provide a more seamless experience for local and customized users.

In this work, we highlight future research along three axes: (i) **cultural sensitivity understanding**, to characterize how state-of-the-art LALMs encode and express culture-specific affect; (ii) **benchmarks**, to provide balanced training/evaluation corpora and protocols that stress cross-cultural generalization; and (iii) **measurement**, to establish metrics that go beyond accuracy and capture fairness, calibration, appropriateness, and interactive quality.

1) Cultural sensitivity understanding: LALMs inherit cultural signals from both text and speech corpora, yet there is limited evidence on *how* these signals manifest in affective perception and generation. We propose a structured agenda:

- **Audit with culture-controlled probes:** construct minimal paired stimuli that vary one cultural dimension at a time (e.g., backchannel tokens, honorifics, irony markers, code-switching, culturally normative laughter) while keeping lexical content constant. Compare recognition and generation outputs across language/dialect groups.
- **Contrastive causal tests:** apply prosodic perturbations (F0 range, speaking rate, intensity contours) and pragmatic rewrites (direct vs. indirect requests) to estimate sensitivities and disentangle lexical, prosodic, and contextual cues.
- **Context infusion studies:** quantify gains from providing sociolinguistic metadata (speaker relation, formality, setting) at inference time via prompts or adapters; ablate each factor to reveal which contexts the model exploits.
- **Human-in-the-loop elicitation:** engage regional experts to curate culturally salient scenarios, iteratively refine prompts/decoding constraints, and annotate failure modes, closing the loop with targeted fine-tuning.

2) Benchmark: Progress requires evaluation settings that reflect real cross-cultural use. We outline benchmark principles and potential tasks:

- **Balanced coverage:** multiple languages/dialects, age groups, genders, regions, interaction types (monologue, dyadic, group), and channels (studio, telephony, far-field).
- **Ontology harmonization:** provide a shared label space (categorical and dimensional) with culture-specific mapping tables; include rationales and display-rule guidance.
- **Rich metadata:** document sociopragmatic context (formality, relationship, setting), elicitation protocol, and annotator background; release dataset/model cards to standardize reporting.
- **Diverse downstream affective computation tasks:** closed-form, open-form SER, reasoning-centric tracks (e.g., reasoning before delivering prediction), and generation-centric tracks (e.g., emotional TTS/voice conversion with parallel/non-parallel targets).

3) Measurement: Traditional accuracy-centric metrics are insufficient for culturally sensitive affect and generative LALMs. We recommend a multi-faceted suite:

- **Core performance:** preserve traditional accuracy-centric scores for classification tasks and correlation-related scores for regression tasks.
- **Robustness:** within the same culture, conduct robust tests under noise, channel, and code-switching; report performance drop relative to clean matched conditions.
- **Open-form mapping quality:** agreement between canonicalized predictions and gold labels (macro F1) plus sensitivity analyses over verbalizer/mapping choices; rationale quality via rubric scores and inter-annotator agreement.
- **Rationale quality (human-in-the-loop):** rate *specificity*, *cultural appropriateness*, *consistency with evidence*, and *coherence* on Likert scales using culturally matched annotators.
- **Counterfactual validity:** when the model predicts that a prosodic change would flip an emotion, apply the change and test whether the prediction/rationale update is consistent.
- **Perceptual quality:** MOS/CMOS for naturalness; affect strength/appropriateness rated by culturally matched listeners and by held-out SER classifiers.

V. Conclusions

LALMs unify recognition and generation for affective speech, yet internationalization remains the central obstacle: vocal affect is culturally grounded, and models trained on imbalanced data misinterpret under-represented users. We advocate a focused agenda: (i) *cultural sensitivity investigation* via systematic audits, causal perturbations, and expert-in-the-loop studies; (ii) *benchmarks* with balanced, well-documented corpora, harmonized ontologies, and protocols spanning closed/open-form SER, reasoning, and generation; and (iii) *measurement* beyond accuracy to include robustness and reasoning-aware evaluation (canonicalization and rationale quality), complemented by perceptual and dialogue metrics for generation/duplex use.

Pursuing this agenda through shared resources, preregistered evaluation harnesses, and human–AI co-design offers a practical path to culturally adapted LALMs that are reliable, appropriate, and inclusive across languages and regions.

Acknowledgment

This work was supported by the National Science and Technology Council(NSTC) under Grant 114-2917-I-564 -022.

REFERENCES

[1] Y. Ma, Y. Zhang, D. Fu, S. Zubicueta Portales, D. Kragic, and M. Fjeld, "Advancing user-voice interaction: Exploring emotion-aware voice assistants through a role-swapping approach," in *International Conference on Human-Computer Interaction*, Springer, 2025, pp. 303–320.

[2] K. K. Coker and R. Thakur, "Alexa, may i adopt you? the role of voice assistant empathy and user-perceived risk in customer service delivery," *Journal of Services Marketing*, vol. 38, no. 3, pp. 301–311, 2024.

[3] A. Votintseva, R. Johnson, and I. Villa, "Emotionally intelligent conversational user interfaces: Bridging empathy and technology in human-computer interaction," in *International Conference on Human-Computer Interaction*, Springer, 2024, pp. 404–422.

[4] T. Deschamps-Berger, "Emotion recognition in emergency call centers: The challenge of real-life emotions," in *2021 9th international conference on affective computing and intelligent interaction workshops and demos (ACIIW)*, IEEE, 2021, pp. 1–5.

[5] N. Shanthi, A. A. Stonier, A. Sherine, *et al.*, "An integrated approach for mental health assessment using emotion analysis and scales," *Healthcare Technology Letters*, vol. 12, no. 1, e12040, 2025.

[6] A. A. Abdelhamid, "Speech emotions recognition for online education," *Fusion: Practice and Applications*, vol. 10, no. 1, pp. 78–87, 2023.

[7] Y. Chu, J. Xu, Q. Yang, *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[8] A. S. Cowen, J. A. Brooks, G. Prasad, *et al.*, "How emotion is experienced and expressed in multiple cultures: A large-scale experiment across north america, europe, and japan," *Frontiers in Psychology*, vol. 15, p. 1 350 631, 2024.

[9] J. A. Brooks, L. Kim, M. Opara, *et al.*, "Deep learning reveals what facial expressions mean to people in different cultures," *Iscience*, vol. 27, no. 3, 2024.

[10] T. A. Nguyen, B. Muller, B. Yu, *et al.*, "Spirit-lm: Interleaved spoken and written language model," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 30–52, 2025.

[11] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[12] C. Wang, M. Rivière, A. Lee, *et al.*, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL 2021-59th Annual Meeting of the Association for Computational Linguistics*, 2021.

[13] N. Das, S. Dingliwal, S. Ronanki, *et al.*, "Speechverse: A large-scale generalizable audio language model," *arXiv preprint arXiv:2405.08295*, 2024.

[14] C. Tang, W. Yu, G. Sun, *et al.*, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[15] P. Anastassiou, J. Chen, J. Chen, *et al.*, "Seed-tts: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024.

[16] M. Abdin, J. Aneja, H. Behl, *et al.*, "Phi-4 technical report," *arXiv preprint arXiv:2412.08905*, 2024.

[17] J. Ye, X.-C. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[18] S. Shen, Y. Gao, F. Liu, H. Wang, and A. Zhou, "Emotion neural transducer for fine-grained speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 10 111–10 115.

[19] D. Chandola, E. Altarawneh, M. Jenkin, and M. Papagelis, "Serc-gcn: Speech emotion recognition in conversation using graph convolutional networks," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 76–80.

[20] L. Martinez-Lucas, W.-C. Lin, and C. Busso, "Analyzing continuous-time and sentence-level annotations for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1754–1768, 2024.

[21] Z. Yang, X. Li, Y. Cheng, T. Zhang, and X. Wang, "Emotion recognition in conversation based on a dynamic complementary graph convolutional network," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1567–1579, 2024.

[22] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[23] T. Qi, S. Wang, C. Lu, Y. Zhao, Y. Zong, and W. Zheng, "Towards realistic emotional voice conversion using controllable emotional intensity," *arXiv preprint arXiv:2407.14800*, 2024.

[24] T. Qi, W. Zheng, C. Lu, Y. Zong, and H. Lian, "Pavits: Exploring prosody-aware vits for end-to-end emotional voice conversion," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 697–12 701.

[25] H.-S. Oh, S.-H. Lee, D.-H. Cho, and S.-W. Lee, "Durflex-evc: Duration-flexible emotional voice conversion with parallel generation," *arXiv preprint arXiv:2401.08095*, 2024.

[26] S. Wang, T. Qi, C. Lu, Z. Luo, and W. Zheng, "Enhancing zero-shot emotional voice conversion via speaker adaptation and duration prediction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[27] Z. Du, J. Lu, K. Zhou, L. Kaushik, and B. Sisman, "Converting anyone's voice: End-to-end expressive voice conversion with a conditional diffusion model," *arXiv preprint arXiv:2405.01730*, 2024.

[28] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audio-visual learning for emotion recognition," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, vol. 16, no. 1, 2025.

[29] J. He, X. Shi, X. Li, and T. Toda, "Mf-aed-aec: Speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 066–11 070.

[30] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6897–6901.

[31] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 4275–4279.

[32] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1912–1926, 2022.

[33] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "Emonet: A transfer learning framework for multi-corpus

speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1472–1487, 2021.
[34] P. Mote, B. Sisman, and C. Busso, "Unsupervised domain adaptation for speech emotion recognition using k-nearest neighbors voice conversion," in *Proc. Interspeech 2024*, 2024, pp. 1045–1049.
[35] B.-H. Su and C.-C. Lee, "Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-gan," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1991–2004, 2022.
[36] K. Chawla, R. Clever, J. Ramirez, G. M. Lucas, and J. Gratch, "Towards emotion-aware agents for improved user satisfaction and partner perception in negotiation dialogues," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 433–444, 2023.
[37] T. Mubassira, M. Hasan, J. Noor, and A. A. A. Islam, "Enhancing emobot: An in-depth analysis of user expectation and satisfaction in an emotion-aware chatbot," in *Proceedings of the 11th International Conference on Networking, Systems, and Security*, 2024, pp. 65–71.
[38] J. Hu, Y. Huang, X. Hu, and Y. Xu, "The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 17–30, 2022.
[39] A. Goel, S. Ghosh, J. Kim, *et al.*, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *arXiv preprint arXiv:2507.08128*, 2025.
[40] P. Wang, S. Lu, Y. Tang, S. Yan, W. Xia, and Y. Xiong, "A full-duplex speech dialogue scheme based on large language model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 13 372–13 403, 2024.
[41] K. Hu, E. Hosseini-Asl, C. Chen, *et al.*, "Efficient and direct duplex modeling for speech-to-speech language model," *arXiv preprint arXiv:2505.15670*, 2025.
[42] Z. Ma, W. Wu, Z. Zheng, *et al.*, "Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 146–11 150.
[43] T. Feng and S. Narayanan, "Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 116–12 120.
[44] J. Santoso, K. Ishizuka, and T. Hashimoto, "Large language model-based emotional speech annotation using context and acoustic feature for speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 026–11 030.
[45] S. Bo-Hao, S. G. Upadhyay, and L. Chi-Chun, "Toward zero-shot speech emotion recognition using llms in the absence of target data," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.
[46] S. G. Upadhyay, W.-S. Chien, B.-H. Su, *et al.*, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2023, pp. 1–8.
[47] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "Lssed: A large-scale dataset and benchmark for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 641–645.
[48] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
[49] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
[50] 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.
[51] Z. Lian, H. Sun, L. Sun, *et al.*, "Ov-mer: Towards open-vocabulary multimodal emotion recognition," *arXiv preprint arXiv:2410.01495*, 2024.
[52] H. Song, J. S. Chan, and C. Ryan, "Differences and similarities in the use of nine emotion regulation strategies in western and east-asian cultures: Systematic review and meta-analysis," *Journal of Cross-Cultural Psychology*, vol. 55, no. 8, pp. 865–885, 2024.
[53] J. L. Tsai, "Ideal affect: Cultural causes and behavioral consequences," *Perspectives on psychological science*, vol. 2, no. 3, pp. 242–259, 2007.
[54] J. L. Tsai, B. Knutson, and H. H. Fung, "Cultural variation in affect valuation.," *Journal of personality and social psychology*, vol. 90, no. 2, p. 288, 2006.