Literature Review

Course Title: Machine Learning

Course Code: CSE 465

Group No: 02

Name	ID
Tasmiah Binte Azad	21225103457
Humayra Kabir Hride	21225103501
Md Ashikur Rahman	21225103471

Speech Emotion Recognition Analysis Using Deep Learning

Research in speech emotion recognition (SER) has evolved through diverse datasets, architectures, and evaluation strategies, yet challenges remain in scalability, generalization, and real-world applicability. Shaik Abdul Khalandar Basha [1] employed CNN-LSTM with attention on RAVDESS, CREMA-D, and TESS, attaining 87.08% accuracy but facing dataset size and real-time deployment limitations. Similarly, Cho and Pappagari [2] applied LSTM with multi-resolution CNN on IEMOCAP, achieving a 21% relative improvement though gains over single acoustic models were modest at 3.4% due to limited generalization. Expanding linguistic diversity, Sultana [3] introduced the SUBESCO Bangla corpus, reporting recognition rates above 70% with anger achieving the highest accuracy (78.3%) and fear the lowest (67.2%), while highlighting issues of ground truth validation and gender bias. In a different domain, Hajdú-Szücs et al. [4] developed a squash evaluation system combining Gaussian models, TDOA localization, and feed-forward networks, yielding 98% accuracy in controlled conditions but dropping to 88% in live matches where noise and rare events hindered performance. Beyond computational models, Schwartz and Pell [5] explored cross-modal priming of prosody and semantics, showing prosody dominance with up to 99% accuracy, though no practical recognition framework was presented. On the modeling front, Kutlimuratov and Cho [6] proposed a CNN-LSTM with attention that achieved near-perfect accuracy on TESS (99.8%) and strong results on RAVDESS (95.7%), yet its robustness in noisy, real-world contexts remained uncertain. Advancing hybrid designs, Hasan et al. [7] introduced EmoFormer, a CNN-Transformer model trained on the META EARS dataset, reaching 90% accuracy for five emotion classes but declining to 83% and 70% for larger emotion sets, underscoring scalability challenges. Complementarily, Pan and Wu [8] developed a CLDNN framework integrating 1D CNN, LSTM, and DNN with data augmentation, reporting 91.87% accuracy on RAVDESS and 88.22% on EMO-DB, with noise injection and pitch-shifting shown to enhance generalization though class imbalance persisted. Finally, Ouyang [9] applied CNN-LSTM with MFCC features on a merged SAVEE and RAVDESS dataset, yielding 61.07% accuracy overall, with anger (75.31%) and neutral (71.70%) recognized most effectively, while disgust lagged (38.33%) due to overlapping negative emotions. Collectively, these studies demonstrate the promise of hybrid deep learning and augmentation strategies in SER, while highlighting persistent limitations in dataset diversity, scalability to many emotion classes, noise robustness, and cross-domain generalization.

Literature review comparison table

Author	**Dataset**	**Research Methods**	**Proposed Architecture**	**Results**	**Challenges/Research gaps**
[1] Shaik Abdul Khalandar Basha	RAVDESS, CREMA-D, TESS	EAS framework,MFCC, STFT, Mel and MLP, SVM, and LSTM classifiers	CNN-LSTM with an attention mechanism for feature selection.	79%-87.08%	Speech Emotion Recognition needs larger datasets, multimodal methods, transfer learning, and real-time readiness for better emotion understanding.
[2] Jaejin Cho1 and Raghavendra Pappagari1	USC-IEMOCAP dataset	LSTM , Multi-resolution CNN analyzes	Acoustic branch (LSTM), Text branch (Multi-resolution CNN) and Fusion + Loss Optimization	21% relative improvement over single acoustic system and 12% improvement over multi-resolution CNN, Only 3.4% improvemen	Speech Emotion Recognition challenges include multimodal fusion, data imbalance with poor generalization, and limited use of verification loss and contextual information.

				t over single acoustic system	
[3] Sadia Sultana	SUBESCO dataset	Corpus Development, human Evaluation, statistical Analysis	Data collection → human evaluation → statistical validation → recognition outcome	Anger , Highest unbiased hit rate: 78.3% and Overall recognition rate above 70%,Fear , Lowest unbiased hit rate: 67.2%.	Data Collection Challenge, Lack of Ground Truth, Weak Emotional Expressions and Gender Effects
[4] Katalin Hajdú-Szücs, Nóra Fenyvesi, József Stéger, and Gábor Vattay	Audio 1 dataset – Controlled training data Audio 2 dataset – Real match data	Gaussian detection, Windowed surprise, TDOA, Gradient descent, FFNN	Gaussian model for detection. TDOA for localization. Feed-forward NN for classification	88% (glass/floor hits) -98% (Front wall)	Strong pipeline, but depends on controlled setup; real match noise may affect accuracy.
[5] Rachel Schwartz and Marc D. Pell	2 Datasets. Prime Stimuli Dataset & Target Stimuli Dataset	A cross-modal priming task (FADT) was used with two conditions: pre-semantic and post-semantic	Cross-modal priming framework. Pre-semantic (prosody only). Post-semantic (prosody and semantics)	87.4% for angry faces to 99% for happy faces	Semantic cues add little benefit; Provides insights on prosody–semantics interaction, but no ML approach or generalizable dataset in recognizing rapid/negative emotions.
[6] Alpamis Kutlimuratov and Young-Im Cho	TESS (Toronto Emotional Speech Set) RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)	ZCR, RMS, MFCCs; preprocessing (normalization, trimming, noise reduction); augmentation (time stretch, pitch shift, noise add)	Hybrid CNN-LSTM model with attention mechanism. CNN for feature extraction LSTM for temporal dependencies	95.7% on RAVDESS dataset to 99.8% on TESS dataset	High accuracy from hybrid model + augmentation; lacks validation on noisy real-world data
[7].	EARS dataset	MFCCs,x-vectors	EmoFormer, a hybrid model	39% - 90%	Performance decreased with an

Rashedul Hasan,			combining CNNs with Transformer encoders for emotion recognition from speech data		increasing number of emotion categories, indicating limitations in scalability
[8] Shing-Tai Pan	RAVDESS, EMODB, IEMOCAP datasets	Pitch shift, time stretch, MFCCs	Two models for speech emotion recognition: 1D CNN-DNN and 1D CLDNN models	63.41% - 95.52%	The collection of speech emotion databases is limited, often recorded by actors, affecting variability in signals
[9] Qianhe Ouyang1	SAVEE, RAVDESS datasets	MFCCs, Librosa package	a CNN-LSTM architecture for extracting features from vocal audio data	61.07%- 75.31%	Overlapping features among emotions complicate recognition in speech emotion detection

References

[1] S. A. K. Basha, “Exploring Deep Learning Methods for Audio Speech Emotion Detection: An Ensemble MFCCs, CNNs and LSTM,” *Applied Mathematics & Information Sciences*, vol. 19, no. 1, Jan. 2025. [Online]. Available: https://dx.doi.org/10.18576/amis/190107

[2] J. Cho and R. Pappagari, “Deep neural networks for emotion recognition combining audio and transcripts,” *arXiv preprint*, arXiv:1911.00432, Nov. 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1911.00432

[3] S. Sultana, “SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla,” *PLOS ONE*, Apr. 30, 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0250173

[4] S. Schwartz and M. D. Pell, “Emotional Speech Processing at the Intersection of Prosody and Semantics,” *PLOS ONE*, Oct. 31, 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0047279

[5] P. Hajdú-Szücs et al., “Audio-based performance evaluation of squash players,” *PLOS ONE*, Apr. 22, 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0194394.g001

[6] B. Kutlimuratov and J. Cho, “Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition,” *Applied Sciences*, vol. 14, no. 23, p. 11342, Dec. 5, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/23/11342

[7] R. Hasan, M. Nigar, N. Mamun, and S. Paul, “EmoFormer: A Text-Independent Speech Emotion Recognition using Hybrid Transformer-CNN model,” in *Proc. 2024 27th Int. Conf. Computer and Information Technology (ICCIT)*, Cox’s Bazar, Bangladesh, Dec. 20–22, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11022032

[8] S.-T. Pan and H.-J. Wu, “Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation,” *Electronics*, vol. 12, no. 11, p. 2436, May 27, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/11/2436

[9] Q. Ouyang, “Speech Emotion Detection Based on MFCC and CNN-LSTM Architecture,” *arXiv preprint*, arXiv:2501.10666, Jan. 18, 2025. [Online]. Available: https://arxiv.org/abs/2501.10666