JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# A Survey on Multimodal Speech Emotion Recognition and Sentiment Analysis: Integrating Acoustic and Linguistic Intelligence for Emotional Understanding

Saksham Lambe
Computer Engineering
Vidyalankar Polytechnic
Vadala,Mumbai
saksham.lambe@vpt.edu.in

Yash Sahu
Computer Engineering
Vidyalankar Polytechnic
Vadala,Mumbai
yash.sahu@vpt.edu.in

Krishna Pawar
Computer Engineering
Vidyalankar Polytechnic
Vadala,Mumbai
krishna.pawar@vpt.edu.in

Anupam Mishra
Computer Engineering
Vidyalankar Polytechnic
Vadala,Mumbai
anupam.mishra@vpt.edu.in

Dr.Vaishali Malkar
Computer Engineering
Vidyalankar Polytechnic
Vadala,Mumbai
vaishali.malkar.vpt.edu.in

***Abstract* —** Emotion Recognition in speech is essential to applications such as human-computer interaction, mental health assessment and call center monitoring. Nearly all the current Speech Emotion Recognition (SER) system concentrates on acoustic dimensions, such as pitch, energy, spectral characteristics but they ignore the content of the spoken words. In this survey, we discuss SER methods, sentiment analysis approaches and emerging multimodal frameworks that correlate between vocal emotion and textual sentiment. We review classical machine learning, deep learning architectures (CNNs, RNNs, BLSTMs, Transformers and attention mechanisms), end-to-end learning and multimodal fusion methods. Key research gaps – lack of multimodal integration, limited cross-lingual generalization, and paucity of production-ready systems are highlighted to pave the way for future scalable emotionally intelligent systems.

## I. INTRODUCTION

Emotion recognition from speech is a field that combines emotion-aware computing, voice signal analysis, and understanding spoken language to identify human emotions. Its applications span mental health assessment, customer service automation, educational technology, and human-computer interaction. Speech simultaneously encodes emotion through acoustic properties – pitch, rhythm, intensity – and through the meaning of words, so a complete picture requires analyzing both channels. In the early days of speech recognition for audio content, researchers would use manually derived acoustic features with traditional classifiers (HMMs, GMMs, SVMs). These worked well on artificial datasets, however they failed in realistic environments because of noise sensitivity, speaker variations and feature representability. The rise of deep learning and its various architectures (CNNs, RNNs, LSTMs, Transformers and attention) brought about automated hierarchical features extraction and performance boost on the scale that was never seen before. However, most of these systems remain unimodal, operating only on acoustic information. In parallel, textual sentiment analysis has matured as a powerful Natural Language Processing (NPL) tool, yet text alone cannot capture vocal cues such as emotional intensity or sarcasm. The resulting gap motivates the present survey, which provides a unified view of SER, sentiment analysis, and multimodal fusion research, identifies unresolved problems, and charts a course for future integrated systems.

## II. SPEECH EMOTION RECOGNITION: SYSTEM COMPONENTS

### A. Emotion Representation Models

Categorical models conceptualize emotion in terms of discrete categories – usually Ekman's six basic emotions and neutral – and naturally corresponds to classification. Dimensional models of emotion represent emotion along continuums of valence, arousal, and dominate, which are more nuanced but come at the cost of requiring regression. The decision depends on applications: interactive systems prefer categorical, subtle analysis is based on dimensional [1].

### B. Benchmark Databases

High-quality, annotated corpora are indispensable. Table I lists the six most popular benchmarks. Acted databases (EmoDB, RAVDESS, TESS) feature clean conditions and right classes, but they may not be representative of real expression. Spontaneous corpora, such as IEMOCAP have annotation difficulties and class imbalance.

**TABLE I: KEY SPEECH EMOTION DATABASES**

| Database | Lang. | Speakers | Utterances | Emotions | Type | Nature | Rate |
|---|---|---|---|---|---|---|---|
| IEMOCAP | EN | 10 (5M/5F) | 10,039 | 9+dims | Categorical | Acted/Improv. | 16 kHz |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| EmoDB | DE | 10 (5M/5F) | 535 | 7 | Categorical | Acted | 16 kHz |
| RAVDESS | EN | 24 (12M/12F) | 7,356 | 8 | Categorical | Acted | 48 kHz |
| SAVEE | EN | 4 (4M) | 480 | 7 | Categorical | Acted | 44.1 kHz |
| CREMA-D | EN | 91 (48M/43F) | 7,442 | 6 | Categorical | Acted | 16 kHz |
| TESS | EN | 2 (0M/2F) | 2,800 | 7 | Categorical | Acted | 24.4 kHz |

C. Preprocessing and Feature Extraction

The raw speech needs to be resampled (usually 16-22 kHz), pre-emphasized filter, framed by a Hamming Window (20-40 ms length, 10-20 ms shift), with VAD amplitudes normalized and possibly denoising. Data augmentation such as adding of Gaussian noise, pitch shifting and time stretching [2] or speed perturbation is commonly used to address the class imbalance and enhance generalization [4].

Manual features are prosodic (F0, energy, duration, Harmonics-to-Noise Ratio), spectral (MFCCs, LPCCs and Chroma) as well as energy statistics (ZCR and RMSE). Standarized corpora, such as IS09 (384 features), IS10 (1,582 featues), allow for intersession comparison from one study to the other. Also, deep learning methods can directly learn features from raw waveforms or log-Mel spectrograms without the requirement of hand-crafted design [1,5].

## III. REVIEW OF SER APPROACHES

### *A. Traditional Machine Learning*

Classical SER relied on SVMs, which work well with high-dimensional features, GMMs, which are used for probabilistic class modeling, and HMMs, which have an explicit temporal structure. Despite careful feature engineering, these methods faced challenges. They struggled with the curse of dimensionality, required dataset-specific feature selection, and could not effectively model complex temporal dependencies.

### *B. Deep Neural Networks and Recurrent Architectures*

Han et al. [3] developed a two-stage DNN–Extreme Learning Machine model in which segment-level DNNs produced emotion probability distributions that were then aggregated by an ELM and achieved 20% (20% improvement over SVM baseline) relative to SVM's on IEMOCAP (UA=57.91% to 63.89%). Lee and Tashev [1] developed an additional model using Bi-LSTM's and attention-weighted pooling to capture bidirectional temporal context that improved the results achieved on IEMOCAP.

### *C. Convolutional Architectures*

The CNNs can also receive log-Mel spectrograms as two-dimensional images, capturing the local temporal/spectro pattern. The authors tested it and found at least a certain amount of success using pyramidal pooling with soft annotator-distribution labels [4] when compared to using single-scale pooling with hard labels. In the study by Mountzouris et al., adding a CNN-related attention layer helped reach 74% accuracy on the SAVEE database and 77% accuracy on the RAVDESS database [7].

### *D. Hybrid CNN–RNN Models*

CNN-BiLSTM ensemble models utilize the speed and efficiency of convolutional neural networks (CNN) to extract local features from raw images and can also process temporal data over long periods of time using long short-term memory networks (LSTM). Chowdhury et al. [2] used lightweight hybrid CNN-BiLSTM ensembles with handcrafted features and found that by optimizing them using adaptive learning rate schedules, dropout, batch normalisation, and early stopping, they were able to achieve near-perfect classification rates: 100% on TESS, 97.57% on the RAVDESS, and between 98.43% and 98.66% on the SAVEE, CREMA-D, and EmoDB datasets.

### *E. Transformer-Based Approaches*

Global dependencies can be modelled by Transformers using self-attention allowing the model to process all positions simultaneously without a sequential bottleneck like a recurrent neural network (RNN). To enhance the modelling capabilities of Transformers, Tang et al. [6] proposed a novel convolutional neural network (CNN)-Transformer architecture that combines lightweight convolution (LCT) blocks, depthwise separable convolutions and coordinate-attention enhanced multi-head self-attention. This architecture produced state of the art performance for the IEMOCAP dataset (72.72% UA, 71.64% WA) and EmoDB dataset (89.51% UA / 90.65% WA) at the time of this publication.

### *F. Pre-trained and Foundation Models*

Large amounts of unlabeled speech data can be used to train self-supervised pre-trained models like wav2vec 2.0, which can then learn powerful acoustic representations that transfer well to emotion tasks. The future of SER systems may be driven by joint pre-training on related tasks, according to large-scale multi-task foundation models like SenseVoice [8], which combines speech recognition, emotion detection, and paralinguistic analysis into a single framework.

### *G. What the Evidence Tells Us*

With the prevalence of hand-crafted features and SVMs (Support Vector Machines), accuracy levels on difficult benchmarks such as IEMOCAP" (Interactive Emotion Recognition Using Multiple Channels Applied in Human-Robot Interaction) (around 60-70%) were respectable but had definite limits. When DNNs (Deep Neural Networks) began to be applied, these limits changed. For example, Han, Kim, and Wilks demonstrated that simply allowing a neural network to develop its own representations for input data improved unweighted accuracy levels on IEMOCAP from approximately 58% to 64%. Although this incremental improvement may not seem like much, it represents a paradigm shift in how we think about emotional perception.

There were benefits to both recurrent and convolutional architectures for emotion detection. LSTM networks were very effective at capturing how emotions evolve over time (i.e. anger builds slowly, sadness gradually diminishes) while CNN networks excelled at capturing unique, sharp energy peaks in the spectral-temporal domain of an emotion. For example, Chowdhury, Hu, et al. [3,2] demonstrated that CNN-BiLSTM hybrid models can be successful through optimization techniques. Optimizing a simple model produces solutions with near-perfect accuracies (i.e. 100% on TESS dataset; >97% on all other datasets). The takeaway from these situations is not that emotion detection has been solved, but that optimization practices (i.e. learning rate scheduling, regularization, smart data augmentation) are as important as the architecture of the model itself. Attention mechanisms reinforced this point. Instead of treating every moment of speech as equally important, attention enables a model to zoom in on the segments that convey emotion and skip over filler words and pauses. The improvements were also robust for different architectures, although the degree of improvement was sensitive to utterance length and variation.

The biggest step forward was the creation of transformers. As outlined by Tang et al.[6], in six they improved on using a mixture of local feature extraction via CNNs with a global self-attention mechanism via their LCT model that achieved 72.72% accuracy on IEMOCAP and more than 90% accuracy on EmoDB compared to what many people would have thought was impossible ten years ago or so. Nevertheless, once again the apparent improvements in the "easier" acted data sources were less than those in the "more difficult" spontaneous data sources that together give the appearance that there is an increasing standard of "high performance" as each benchmark has the capacity to affect its own measure of performance.

In short, it was not one element that was the success factor, but the outcome of an optimal combination of all the following elements: learnt features + temporal modelling + attention + proper training. Generalization remains the largest difficulty—the model that performs superlatively in a controlled environment will generally fail catastrophically when the recording conditions differ from those of training (i.e. speaker demographics) and/or the model is working with a different emotional style than developed during the training phase. It will be in closing the gap between benchmark performance and real-world robustness that the toughest challenge awaits the community.

## IV. SENTIMENT ANALYSIS AND MULTIMODAL FUSION

### A. Textual Sentiment Analysis

Sentiment Analysis is used to figure out how people feel about something they wrote. It looks at the words to see if they are good or bad or just okay. At first people used lists of words to do this. Then they started using computers to teach themselves how to do it better. Now computers use ways like Naïve Bayes and SVM and even more advanced ways like LSTM and BERT to understand how people feel from what they write. These ways work well when people write things down. They do not work as well when people talk because they cannot hear the way the person is talking. When people talk they use things, like how loud or soft they talk. How fast they talk to show how they feel and Sentiment Analysis has a hard time understanding these things.

### B. Argument for Multimodal Convergence

Human emotions can be described as multimodal. The emotional affect of a given utterance can be defined by both its vocal (how) and content (what) characteristics. For example, a speech can have an acoustic positive (happy) affect while at the same time having a semantic negative (sad) affect. This same phenomenon can be applied to reject surface syntax. By combining all of these modalities, it is possible to reduce ambiguity and increase robustness. In addition to EEG and galvanic skin response signals, facial expressions provide additional modalities that could potentially facilitate future integration of these channels.

### C. Fusion Strategies

Early (or feature-level) fusion involves the concatenation of the acoustic and language vectors prior to a joint classifier, enabling interaction learning but with rather high-dimensional inputs. Late (or decision-level) fusion involves the combination of decisions produced by separately trained unimodal models, and preserves modality-specific optimization. Intermediate fusion involved the combination of internal representations from mid-network layers, trading off flexibility and CPU needs. Cross-modal attention mechanisms have proved to be a sound middle ground, learning which speech units are more informative in the context of text information and vice versa [5].

## V. RESEARCH GAPS AND FUTURE DIRECTIONS

### A. Cross-Corpus Generalization

The systems are trained on one data set but perform poorly on another because of mismatches in the conditions of recording, demographics, and styles of expressing emotions. “Research shows us that annotators inaccurately label the training data, and as a result, our models may not be able to detect emotions in everyday speech,” says Vashisht Dhanraj, Director of Inclusive AI at Microsoft.

### B. Spontaneous vs. Acted Emotion

Models trained on datasets such as EmoDB and RAVDESS often achieve over 95% accuracy, but the same models typically yield much lower performance when predicting emotions from improvised speech samples within the IEMOCAP corpus (70–73%). As a result, the development of more sophisticated modeling approaches is necessary to help capture naturally occurring, subtle forms of emotional expression.

### C. Real-Time Efficiency

To deploy large transformer models and deep CNN-LSTM ensembles to edge devices and mobile platforms where latency is unacceptable for interactive applications it is critical to develop lightweight, hardware-friendly architectures through methods such parameter pruning, weight quantization, and knowledge distillation while maintaining high levels of accuracy.[2]

### D. Multimodal Integration Depth

However, while these approaches are clearly complementary to each other, fully integrated multimodal systems remain largely an elusive dream. In most of the systems, acoustic and linguistic streams are treated separately and only combined at the decision level. Future research should also explore deep cross-modal attention and shared latent spaces, as well as training methods that softly handle missing modality.

### E. Interpretability and Fairness

Deep learning SER models are essentially black boxes. Explainable methods (attention visualization, LIME, SHAP) are required for trust-building as well as systematic error analysis. It is equally important to investigate bias across speaker gender, age, accent, and cultural background, gaps that current benchmark-centric evaluation largely obscures.

## VI. CONCLUSION

This survey has traced the evolution of Speech Emotion Recognition from hand-crafted acoustic features combined with classical classifiers to sophisticated deep learning architectures incorporating CNNs, LSTMs, Transformers, and cross-modal attention. Key milestones include the

DNN–ELM pipeline [3], Bi-LSTM with attention [1], pyramidal CNN training [4], lightweight CNN–BiLSTM ensembles [2], multidimensional CNN-Transformer models [6], and emergent foundation models [8].

The essential point being that acoustic analysis is not sufficient for human level emotional intelligence. The next step are unified multimodal systems based on joint modeling of the vocal and linguistic streams. To enable robust, real-time, interpretable and generalisable emotional intelligence in the wild will entail continuing to make progress on cross-corpus evaluation, spontaneous speech modeling, efficient architecture design and ethical fairness—challenges this survey aims to help galvanise.

## ACKNOWLEDGMENT

We thank the Department of Computer Engineering, Vidyalankar Polytechnic, for institutional support. Special gratitude is extended to Dr. Vaishali Malkar for her invaluable guidance and mentorship throughout this work.

## REFERENCES

[1] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in Proc. Interspeech, Dresden, Germany, Sep. 2015.

[2] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with lightweight deep neural ensemble model using hand-crafted features," Scientific Reports, in press, 2025.

[3] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Proc. Interspeech, Singapore, Sep. 2014.

[4] S. Parthasarathy and I. Tashev, "Convolutional neural network techniques for speech emotion recognition," in Proc. ICASSP, 2016.

[5] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," Electronics, vol. 10, no. 1163, 2021.

[6] X. Tang, Y. Lin, T. Dang, Y. Zhang, and J. Cheng, "Speech emotion recognition via CNN-transformer and multidimensional attention mechanism," Speech Communication, in press, 2025.

[7] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech emotion recognition using convolutional neural networks with attention mechanism," in Proc. IEEE ICTAI, 2021.

[8] K. An et al., "FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs," arXiv:2407.04051, 2024.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.