**Speech Emotion Recognition**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.No.** | **Year** | **Author(s)** | **Tool** | **Technique** | **Result** | **Conclusion/ Future Scope** | **DOI** | **Links** |
| 1 | 2021 | Taiba Majid Wani,  Teddy Surya Gunawan | These systems target the speaker’s existence of varied emotions by extracting and classifying the prominent features from a  preprocessed speech signal | SVM, ANN, CNN | The paper reviews various databases, features, and classifiers used in speech emotion recognition from assorted languages | The paper scrutinizes the  current state of understanding on SER systems and sketches out the prominence of the research gap for  consideration and analysis by other related researchers,  institutions, and regulatory bodies | 10.1109/ACCESS.  2021.3068045 | https://ieeexplore.ieee.  org/abstract/document/938300 0/ |
| 2 | 2018 | Monorama Swain,  Aurobinda Routray &  P. Kabisatpathy | Databases, features and classifiers | The paper reviews various databases, features, and classifiers used in speech emotion recognition from assorted languages | It discusses the identification of emotional content from speech and the development of systems and algorithms for this purpose | The study concludes with insights into the features, databases, and classifiers used from 2000 to 2017 for speech emotion recognition, suggesting a multidisciplinary approach involving psychology, linguistics, and artificial intelligence. | 10.1007/s10772-  018-9491-z | https://link.springer.  com/article/10.1007/s10772018-9491-z |
| 3 | 2019 | Mehmet Berkehan Akc¸ay, Kaya Oguz | Autoencoders, Multitask learning, Attention  Mechanism, Transfer Learning, Adversarial technique, | HMM, CNN, kNN, MLR, Support Vector Regression | Current SER systems struggle with cross-language and multi-speech challenges, needing improved algorithms  for better handling of intonation differences and speech separation. | The future scope of SER systems includes the application of more powerful hardware and deep learning algorithms for improved recognition rates, leading to wider usage in daily life and areas like human-computer interaction and healthcare. | 10.1016/j.specom.  2019.12.001 | https://www.sciencedirect.  com/science/article/pii/S01676  39319302262 |
| 4 | 2014 | Huang, Z., Dong, M., Mao, Q., & Zhan, Y. | Speech Emotion Recognition | CNN (Convolutional Neural Network) | The study likely discusses the outcomes of applying CNN to speech emotion recognition. | The paper outline conclusions and future research directions in the field of speech emotion recognition using CNN. | 10.1145/2647868.  2654984 | https://dl.acm.org/doi/abs/10. 1145/2647868.2654984 |
| 5 | 2017 | Fayek, H. M., Lech, M., & Cavedon, L. | Neural Network | CNNs | The paper evaluates deep learning architectures for  Speech Emotion Recognition. | The paper studies the future of neural nets in modern signal processing | 10.1016/j.neunet.  2017.02.013 | https://www.sciencedirect.  com/science/article/pii/S08936  0801730059X |
| 6 | 2020 | Manas Jain, Shruthi Narayan, Pratibha  Balaji, Bharath K P,  Abhijit Bhowmick,  Karthik R, Rajesh  Kumar Muthu | Support Vector Machines | SVM, LPCC, MFCC | The overall accuracy of the system was found to be 85.085%. | The study concluded that gender-dependent classifier shows better accuracy than OAA classifier algorithm, and future work will focus on combining other feature values with MFCC and MEDC to increase system accuracy | --- | https://arxiv.org/abs/2002. 07590 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 7 | 2014 | Han, Kun and Yu,  Dong and Tashev, Ivan | Utilization of Deep Neural Networks (DNNs) to extract high-level features from raw data for speech emotion recognition | Extreme Learning Machine (ELM), a single-hidden-layer neural network, used for utterance-level emotion classification. | The approach led to a 20%  relative accuracy improvement compared to state-of-the-art methods. | The paper suggests that neural networks are promising for learning emotional information from low-level acoustic features and indicates potential for future improvements in speech emotion recognition. |  | https://www.microsoft.com/enus/research/publication/speech -emotion-recognition-usingdeep-neural-network-andextreme-learning-machine/ |
| 8 | 2020 | Issa, D., Fatih Demirci, M., & Yazici, A. | Deep Convolution Networks | DNNs, DCNs | the classification accuracy of Model E is 95.71%. | The proposed onedimensional deep CNN framework with a combination of five audio features demonstrates superior performance on SER tasks, particularly in accuracy, simplicity, and generality, though further research on feature inclusion, | 10.1016/j.bspc.  2020.101894 | https://www.sciencedirect.com/science/article/pii/S1746809420300501 |
| 9 | 2015 | Kunxia Wang, Ning An,  Bing Nan Li, Yanyong  Zhang, & Lian Li | The paper discusses Speech Emotion Recognition. | The study utilizes Fourier Parameters. | The research achieved advancements in emotion recognition accuracy. | The paper suggests potential  improvements and future directions in the field of affective computing. | 10.1109/TAFFC.  2015.2392101 | https://ieeexplore.ieee.  org/abstract/document/700999 7/ |
| 10 | 2016 | Sayan Ghosh, Eugene  Laksana, LouisPhilippe Morency,  Stefan Scherer | Spectrogram Features: Utilization of spectrogram features extracted from speech and glottal flow signals12.  Stacked Autoencoder: A stacked autoencoder is employed for spectrogram encoding to learn features directly from the signal. | Stacked Autoencoders: Utilized for pretraining to learn lower-dimensional representations of input data1.  BLSTM-RNN: A Bidirectional Long-Short Term Memory Recurrent Neural Network used for sequence classification of emotions. | Emotion Classification Accuracy: The proposed approach achieved competitive results in classifying emotions from speech signals. The combination of stacked autoencoders and BLSTMRNN yielded an accuracy of approximately 70% on the | Conclusions: The study found that features learned from spectrograms are highly discriminative for emotion classification and comparable  to state-of-the-art approaches1. Filtering out speaker and phonetic information by inverse filtering reduces confusion | 10.21437  /Interspeech.2016692 | http://multicomp.cs.cmu.edu/wpcontent/uploads/2017/09/2016\_Interspeech\_Ghosh\_Represe ntation.pdf |
| 11 | 2020 | Alberto Cano | The research utilized Melfrequency cepstrum coefficients (MFCC) and modulation spectral (MS) features extracted from speech signals. | The study involves Automatic Speech Emotion Recognition using machine learning techniques. | The RNN classifier achieved the best accuracy of 94% on the Spanish database without speaker normalization (SN) and with feature selection (FS). All classifiers reached an accuracy of 83% for the Berlin database with SN and FS. | The study suggests that while RNN performs well with more data, SVM and MLR  are more practical for limited data. Future work includes enhancing system robustness by combining databases and classifier fusion, and testing the system in pedagogical | 10.5772  /intechopen.84856 | https://hal.science/hal02432557/ |
| 12 | 2018 | Tzirakis, Panagiotis; Zhang, Jiehao;  Schuller, Bjorn W. | The research utilized a  Convolutional Neural Network (CNN) for feature extraction from raw signals, and a 2-layer Long ShortTerm Memory (LSTM) for contextual information processing. | The study presented an endto-end model for continuous speech emotion recognition,  which significantly  outperformed state-of-the-art methods for the RECOLA database in terms of concordance correlation coefficient. | 78% Arousal | The paper suggests exploring deeper CNN models and larger databases for improved audio analysis using raw signals, with considerations for kernel size and pooling size in model design | 10.1109/ICASSP.  2018.8462677 | https://ieeexplore.ieee.  org/abstract/document/846267 7/ |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13 | 2014 | Mao, Qirong; Dong,  Ming; Huang,  Zhengwei Zhan,  Yongzhao | CNN, SAVEE | The paper introduces a twostage training process involving Stacked AutoEncoders (SAE) and  Supervised Discriminative Feature Analysis (SDFA) with a novel objective function. | 71.8% on SAVEE | Plan to extend the proposed method in this paper and evaluate  its performance on naturalistic speech data | 10.1109/TMM.  2014.2360798 | https://ieeexplore.ieee.  org/abstract/document/691301 3/ |
| 14 | 2017 | Deng, Jun; Xu,  Xinzhou; Zhang,  Zixing; Fruhholz,  Sascha; Schuller,  Bjorn | The paper doesn’t specify the tools used directly, but it discusses the use of semisupervised autoencoders for speech emotion recognition. | The focus is on semisupervised learning with autoencoders, emphasizing the combination of generative and discriminative training. | The methods improved recognition performance by learning from unlabeled data, especially with a small number of labeled examples, leading to state-of-the-art performance. | The research indicates that the model can effectively use both labeled and unlabeled data for speech emotion recognition. Future work may explore residual neural networks and their potential in this field. | 10.1109/TASLP.  2017.2759338 | https://ieeexplore.ieee.  org/abstract/document/805987 2/ |
| 15 | 2015 | Christos-Nikolaos  Anagnostopoulos,  Theodoros Iliou,  Ioannis Giannoukos | The research surveyed emotion recognition methods, focusing on feature extraction and classification techniques, without specifying tools used | The research surveyed emotion recognition methods, focusing on feature extraction and classification techniques, without specifying tools used | The study emphasizes the  lack of uniformity in evaluation methods, making direct comparisons difficult | It suggests that feature sets and classification methods need further exploration and that hybrid classifiers and ensembles could be more effective. The development of a common database for emotion recognition is challenging due to crosscultural diversities. | 10.1007/s10462-  012-9368-5 | https://link.springer.  com/article/10.1007/s10462012-9368-5 |
| 16 | 2015 | Zheng, W. Q.; Yu, J. S.; Zou, Y. X. | deep learning frameworks  and speech processing libraries | The study involves an experimental study of speech emotion recognition based on deep convolutional neural networks. | 40% classification Accuracy | The research opens up possibilities for further advancements in affective computing, particularly in improving the accuracy and efficiency of emotion recognition systems, and expanding their application in various fields such as human-computer interaction | 10.1109/ACII. 2015.7344669 | https://ieeexplore.ieee.  org/abstract/document/734466 9/ |
| 17 | 2016 | Wootaek Lim,  Daeyoung Jang, Taejin  Lee | DNN, CNN, RNN, LSTM | The study utilizes convolutional and recurrent neural networks for speech emotion recognition. | 86.65% | The authors plan to study the audio/video based multimodal emotion recognition task in the future. They also expect that using  more concatenated CNNs can give better results | 10.1109/APSIPA.  2016.7820699 | https://ieeexplore.ieee.  org/abstract/document/782069 9/ |
| 18 | 2015 | Jin, Qin; Li, Chengxin; Chen, Shizhe; Wu, Huimin. | Low-level acoustic feature extraction tools, Gaussian Supervectors, Bag-of-Words (BoW) feature. | The paper presents feature representations from both acoustic and lexical levels, including statistics, acoustic codewords, and a new feature called emotion vector (eVector). | Late fusion of acoustic and lexical features achieved a four-class emotion recognition accuracy of 69.2%. | The study suggests that combining different feature representations can enhance emotion recognition systems. | 10.1109/ICASSP.  2015.7178872 | https://ieeexplore.ieee.  org/abstract/document/717887 2/ |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 19 | 2018 | Liu, Z.-T., Wu, M.,  Cao, W.-H., Mao, J.-  W., Xu, J.-P., & Tan,  G.-Z | Decision Tree | The study involves speech emotion recognition using feature selection and an extreme learning machine decision tree. | The proposed framework achieved an 89.6% recognition rate on average using the Chinese speech database from the Institute of  Automation of the Chinese Academy of Sciences (CASIA). | Future research will explore novel speech emotional features, other feature extraction methods like deep learning, feature selection based on evolutionary computation, and testing the practicability of the method with different speech databases. The approach | 10.1016/j.neucom.  2017.07.050 | https://www.sciencedirect.  com/science/article/pii/S09252  31217313565 |
| 20 | 2020 | Koduru, Anusha;  Valiveti, Hima Bindu; Budati, Anil Kumar | Filters for preprocessing, Mel  Frequency Cepstral  Coefficients (MFCC),  Discrete Wavelet Transform (DWT), pitch, energy, Zero crossing rate (ZCR) algorithms for feature extraction, and Global feature algorithm for feature selection. | The study focused on improving speech emotion recognition using various feature extraction algorithms | The accuracy achieved was 70% for SVM, 85% for Decision Tree, and 65% for LDA. | The proposed system outperformed existing work in accuracy and processing time, making it suitable for all kinds of signals and providing a better speech recognition rate. | 10.1007/s10772-  020-09672-4 | https://link.springer.  com/article/10.1007/s10772020-09672-4 |
| 21 | 2019 | T. Özseven | Forward Feature Selection (FFS), Backward Feature  Selection (BFS), Sequential  Floating Forward Selection  (SFFS), wrapper approach,  Principal Component  Analysis (PCA), Linear Discriminate Analysis (LDA), and Fast Correlation-Based Filter (FCBF). | The study focused on Speech Emotion Recognition (SER) systems, analyzing speech by acoustic methods to obtain features that characterize emotional content efficiently and are independent of the speaker or word content. | The highest success rate mentioned is 93.78% using Support Vector Machine (SVM) in EMO-DB studies for 7 emotions. | The highest success rate mentioned is 93.78% using Support Vector Machine (SVM) in EMO-DB studies for 7 emotions. | 10.1016/j.apacoust.  2018.11.028 | https://www.sciencedirect.  com/science/article/pii/S00036  82X18309915 |
| 22 | 2017 | Zhang, Shiqing;  Zhang, Shilliang;  Huang, Tiejun; Gao,  Wen | Deep Convolutional Neural Networks (DCNN), AlexNet model, ImageNet dataset | The paper utilizes DCNN with log Mel-spectrograms to bridge the affective gap in speech signals. It employs Discriminant Temporal Pyramid Matching (DTPM) for feature aggregation and Support Vector Machines (SVM) for classification | The approach shows promising performance on datasets like EMO-DB, RML,  eNTERFACE05, and BAUM-  1s, with substantial improvement after fine-tuning on target datasets. | The paper suggests further investigation into deep learning for speech emotion recognition and continuous dimension emotion recognition, combining CNN with LSTM for temporal cues. | 10.1109/TMM.  2017.2766843 | https://ieeexplore.ieee.  org/abstract/document/808517 4/ |
| 23 | 2018 | Yoon, Seunghyun;  Byun, Seokhyun; Jung,  Kyomin | RNN | A novel deep dual recurrent encoder model utilizing both text data and audio signals, encoded using dual recurrent neural networks (RNNs). | Achieved accuracies ranging from 68.8% to 71.8% on the IEMOCAP dataset, outperforming previous stateof-the-art methods. | Plans to extend modalities to include video inputs and investigate the application of the attention mechanism to enhance multimodal classification tasks. | 10.1109/SLT.  2018.8639583 | https://ieeexplore.ieee.  org/abstract/document/863958 3/ |
| 24 | 2017 | Aldeneh, Zakaria;  Provost, Emily Mower | Convolutional Neural  Networks (CNNs), Mel  Filterbanks (MFBs) | The study applied CNNs to low-level acoustic features to identify emotionally salient regions without relying on utterance-level statistics. It compared the performance of CNNs with traditional classifiers using handengineered features. | The research demonstrated a significant improvement in emotion recognition accuracy  using regional saliency with CNNs and MFBs, showing better results than networks relying on utterance-level statistics. | he paper suggests CNNs with MFBs as a viable  alternative to traditional  SVMs for Speech Emotion Recognition (SER). Future work includes exploring the combination of filters with varying widths and additional  Low-Level Descriptors | 10.1109/ICASSP.  2017.7952655 | https://ieeexplore.ieee.  org/abstract/document/795265 5/ |

(LLDs).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 25 | 2019 | Khalil, Ruhul Amin;  Jones, Edward; Babar,  Mohammad  Inayatullah; Jan,  Tariqullah; Zafar,  Mohammad Haseeb; Alhussain, Thame | RNN, CNN | The paper reviews Deep  Learning techniques like  DBM, RNN, DBN, CNN, and AE for speech emotion recognition (SER). | The research paper presents the shortcomings of DNNs, authors proposed that the Limitations of deep learning techniques include their large layer-wise internal architecture, less efficiency for temporally-varying input data and over-learning during memorization of layer-wise | It discusses the performance and limitations of these techniques and suggests directions for future SER systems. | 10.1109/ACCESS.  2019.2936124 | https://ieeexplore.ieee.  org/abstract/document/880518 1/ |
| 26 | 2017 | Mirsamadi,  Seyedmahdad;  Barsoum, Emad;  Zhang, Cha | RNN | The study introduces a novel feature pooling strategy using local attention to focus on emotionally salient parts of speech signals. | The proposed solution outperformed existing emotion recognition algorithms in accuracy when evaluated on the IEMOCAP corpus. | The paper suggests that deep RNNs can effectively learn both frame-level and temporal aggregation of speech features for emotion recognition. The attention mechanism used allows for focusing on emotionally significant segments, indicating potential for future | 10.1109/ICASSP.  2017.7952552 | https://ieeexplore.ieee.  org/abstract/document/795255 2/ |
| 27 | 2020 | Mustaqeem, ; Sajjad,  Muhammad; Kwon,  Soonil | The research utilized STFT algorithm, CNN model (specifically Resnet with FC-  1000 layers), and deep BiLSTM. | A novel SER framework was introduced, employing RBFNbased clustering for key sequence selection and feature normalization before processing. | Achieved accuracies of  72.25% on IEMOCAP,  85.57% on EMO-DB, and 77.02% on RAVDESS datasets. | The architecture can be applied to other applications and further explored using DBN, GRU, and spike networks for improved accuracy and reduced complexity. | 10.1109/ACCESS.  2020.2990405 | https://ieeexplore.ieee.  org/abstract/document/907878 9/ |
| 28 | 2017 | Badshah, Abdul Malik;  Ahmad, Jamil; Rahim,  Nasir; Baik, Sung  Wook | The research utilized deep convolutional neural networks (CNNs) for feature learning from spectrograms. | Speech signals were represented as spectrograms, serving as input to the CNN model, which consisted of three convolutional and three fully connected layers. | The first experiment with the Berlin emotions dataset yielded satisfactory results for most emotions except fear. The second experiment involved fine-tuning a pretrained AlexNet model, but the results were not satisfactory. | Further work is needed to improve the framework for effective and robust emotion recognition. Plans include using more data and relatively complex models to enhance SER performance | 10.1109/PlatCon.  2017.7883728 | https://ieeexplore.ieee.  org/abstract/document/788372 8/ |
| 29 | 2019 | Zhao, J., Mao, X., & Chen, L. | The research utilized 1D &  2D CNN LSTM networks for feature learning from speech and log-mel spectrogram | he study investigated learning local correlations and global contextual information from raw audio clips and log-mel spectrograms using Local Feature Learning Blocks (LFLBs) and Long ShortTerm Memory (LSTM) layers. | The 2D CNN LSTM network achieved recognition accuracies of 95.33% and 95.89% on Berlin EmoDB for speaker-dependent and speaker-independent experiments, respectively, and 89.16% and 52.14% on IEMOCAP database. | The paper highlights the need to improve understanding of how the networks recognize emotion and to uncover the “black box” of these networks. | 10.1016/j.bspc.  2018.08.035 | https://www.sciencedirect.  com/science/article/pii/S17468  09418302337 |
| 30 | 2015 | Jinkyu Lee and Ivan Tashev | The research utilizes a  Recurrent Neural Network (RNN) with a Bidirectional Long Short-Term Memory (BLSTM) model. | The paper introduces an efficient learning algorithm for speech emotion recognition that accounts for long-range context and label uncertainty | The proposed system achieved a 12% improvement in weighted accuracy over the baseline DNN-ELM based emotion recognition system | The paper concludes that the RNN-based framework with the new learning approach effectively handles long contextual effects and label uncertainty in speech emotion recognition, offering insights for future research2 |  | https://www.microsoft.com/enus/research/publication/highlevel-feature-representationusing-recurrent-neuralnetwork-for-speech-emotionrecognition/ |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | 2019 | Michael Neumann, Ngoc Thang Vu | Autoencoders (AE),  Convolutional Neural  Network (CNN), t-distributed  Stochastic Neighbor  Embedding (t-SNE) | Unsupervised representation learning on large unlabeled speech corpora using autoencoders to learn suitable features in an unsupervised manner. These representations are then integrated into a CNN-based emotion classifier to improve recognition accuracy. | The integration of representations learned by unsupervised autoencoders into a CNN-based emotion classifier led to improved recognition accuracy in speech emotion recognition tasks. Visualizations of the representations using t-SNE provided insights into what | Conclusion: The study demonstrated the effectiveness of utilizing representation learning on unlabeled speech data to enhance speech emotion recognition systems.  Future Scope: Future research could explore more | 0.1109/ICASSP.  2019.8682541 | https://ieeexplore.ieee.  org/abstract/document/868254 1/ |
| 32 | 2021 | Babak Joze  Abbaschian,Daniel  Sierra-Sos and Adel  Elmaghraby | ANN, CNN, LSTM, attention mechanism, autoencoders, GAN. | Deep learning for emotion recognition from speech. | SER accuracy exceeds 90% in controlled environments. | Deep learning revolutionizes speech emotion recognition. | 10.3390/s21041249 | https://www.mdpi.com/14248220/21/4/1249 |
| 33 | 2016 | George Trigeorgis, Fabien Ringeval,  Raymond Brueckner,  Erik Marchi, Mihalis A.  Nicolaou, Björn  Schuller, Stefanos  Zafeiriou | CNN, LSTM networks. | End-to-end speech emotion recognition. | utperforms traditional  methods on RECOLA database. | CNN-LSTM combination improves emotion recognition. | 10.1109/ICASSP.  2016.7472669 | https://ieeexplore.ieee.  org/abstract/document/747266 9/ |
| 34 | 2019 | Seunghyun Yoon; Seokhyun Byun;  Subhadeep Dey;  Kyomin Jung | BLSTM, multi-hop attention mechanism. | Integration of acoustic and lexical data for emotion classification. | 6.5% relative improvement in weighted accuracy. | Proposed technique outperforms state-of-the-art systems. | 10.1109/ICASSP.  2019.8683483 | https://www.mdpi.com/14248220/20/1/183 |
| 35 | 2019 | Mustqueem, Soonil Kwon | Deep Stride Convolutional  Neural Network (DSCNN)  Short-Term Fourier  Transformation (STFT) Adaptive thresholding technique  SoftMax classifier | Pre-processing to remove noise and silent portions Spectrogram generation for 2D representation  CNN architecture with special strides in convolutional layers Plain nets strategy for feature  learning | Increased accuracy of SER by 7.85% and 4.5% on  IEMOCAP and RAVDESS  datasets respectively Reduction in computational complexity by 34.5 MB | Proposed DSCNN architecture improves SER accuracy while reducing computational complexity. Pre-processing and spectrogram generation enhance feature extraction for emotion recognition. | 10.3390/s20010183 | https://ieeexplore.ieee.  org/abstract/document/868348 3/ |
| 36 | 2017 | Michael Neumann, Ngoc Thang Vu | Attentive Convolutional  Neural Network (CNN)  Multi-view learning objective function | Experimentation with different input signal lengths  Evaluation of various acoustic feature types  Analysis of improvised vs. scripted speech emotion data | Recognition performance depends on speech type, regardless of input features State-of-the-art results achieved on improvised speech data in IEMOCAP | Emotion recognition performance varies based on speech type, influencing system effectiveness. | 10.48550/arXiv.  1706.00612 | https://arxiv.org/abs/1706. 00612 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 37 | 2017 | Abdul Malik Badshah, Nasir Rahim, Noor  Ullah, Jamil Ahmad,  Khan Muhammad, Mi  Young Lee, Soonil  Kwon & Sung Wook  Baik | Deep Convolutional Neural Network (CNN)  Rectangular kernels for feature extraction | Feature extraction from spectrograms  Use of rectangular kernels and max pooling for discriminative features | Outperforms state-of-the-art methods on Emo-DB and Korean speech dataset | Rectangular kernels improve feature extraction from spectrograms, enhancing speech emotion recognition performance. | 10.1007/s11042-  016-4041-7 | https://link.springer.  com/article/10.1007/s11042017-5292-7 |
| 38 | 2018 | Björn W. Schuller | Speech emotion recognition techniques and benchmarks | Comprehensive review of emotion recognition in speech over two decades | Summarizes progress and ongoing trends in speech emotion recognition research | Provides insights for further research and development in emotion recognition from speech. | 10.1145/3129340 | https://dl.acm.org/doi/abs/10. 1145/3129340 |
| 39 | 2018 | Promod Yenigalla,  Abhay Kumar, Suraj  Tripathi, Chirag Singh,  Sibsambhu Kar,  Jithendra Vepa | Phoneme sequences  Spectrograms  Deep neural networks (CNN)  Phoneme embedding  IEMOCAP dataset | Phoneme-based CNN model Spectrogram-based CNN model  Multi-channel CNN model combining phoneme and spectrogram features | Phoneme and spectrogram combined CNN achieved over 4% increase in accuracy.  State-of-the-art results  surpassed on IEMOCAP dataset.  Improved accuracy in emotion recognition, especially for improvised | The combination of spectrogram-based 2D CNN  with phoneme embedding significantly improves emotion recognition accuracy. This approach offers over 4% increase in both overall accuracy and average class accuracy compared to existing | 10.21437 /Interspeech | https://abhayk1201.github.  io/files/paper2.pdf |
| 40 | 2014 | Jun Deng; Zixing  Zhang; Florian Eyben;  Björn Schuller | Adaptive Denoising Autoencoder  Unsupervised domain adaptation  Speech emotion corpora  Machine learning methods | Adaptation of autoencoder for domain adaptation Matching feature space representation for source and target sets | Significant improvement over baseline performance Outperforms related feature domain adaptation methods | Effective method for addressing discrepancy between training and test data  Enhances performance of speech emotion recognition engines. | 10.1109/LSP.  2014.2324759 | https://ieeexplore.ieee.  org/abstract/document/681752 0/ |
| 41 | 2019 | Hao Meng; Tianhao  Yan; Fei Yuan;  Hongwei Wei | Dilated CNN  Residual block  BiLSTM  Attention mechanism  Center loss | 3D Log-Mel spectrograms as input  Dilated CNN with residual block  BiLSTM for long-term dependencies  Attention mechanism for feature extraction  Improved loss function with center loss | 74.96% unweighted accuracy  (speaker-dependent) in  IEMOCAP  69.32% unweighted accuracy  (speaker-independent) in  IEMOCAP  90.78% accuracy (speakerdependent) in Berlin EMODB 85.39% accuracy (speakerindependent) in Berlin | Effective emotion recognition from speech signals Better performance compared to previous stateof-the-art methods. | 10.1109/ACCESS.  2019.2938007 | https://ieeexplore.ieee.  org/abstract/document/881791 3/ |
| 42 | 2018 | Mingyi Chen; Xuanji  He; Jing Yang; Han  Zhang | Three-dimensional attentionbased convolutional  recurrent neural networks  (3D-ACRNN)  Mel-spectrogram with deltas and delta-deltas | Utilization of deltas and deltadeltas for personalized features  Three-dimensional attention mechanism for learning discriminative features Input: Mel-spectrogram with deltas and delta-deltas | State-of-the-art performance in unweighted average recall Effective reduction of misclassification  Tested on IEMOCAP and  Emo-DB corpus | Demonstrated effectiveness in SER  Mitigated influence of emotionally irrelevant factors | 10.1109/LSP.  2018.2860246 | https://ieeexplore.ieee.  org/abstract/document/842102 3/ |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 43 | 2019 | Yuanchao Li  , Tianyu Zhao  , Tatsuya Kawahara | End-to-end (E2E) multitask learning  Self-attention mechanism  CNN-BLSTM architecture | Direct feature extraction from spectrogram  Self-attentional CNN-BLSTM model for salient feature extraction  Multitask learning with gender classification as auxiliary task | Outperformed state-of-the-art by 7.7% absolute improvement  Enhanced overall accuracy by integrating gender classification  Achieved significant improvement in emotion classification on IEMOCAP | Proposed model effectively  handles variability in speech and emotion  Offers potential for real-world applications in human-robot interaction  Future work includes incorporating additional paralinguistic tasks | 10.21437  /Interspeech.20192594 | https://www.isca-archive. org/interspeech\_2019/li19n\_int erspeech.pdf |
| 44 | 2019 | Noushin Hajarolasvadi and Hasan Demirel | 3D Convolutional Neural  Networks (CNN)  K-means clustering  Spectrograms | Frame splitting and feature extraction  K-means clustering for selecting keyframes  Training 3D CNN with 10-fold cross-validation | Superior performance  compared to state-of-the-art methods  Outperformed pre-trained 2D  CNNs  Tested on multiple databases: SAVEE, RML, eNTERFACE’05 | Selected keyframes enhance performance of 3D CNN Future work: compare with pre-trained 3D architectures, explore multimodal fusion. | 10.3390/e21050479 | https://www.mdpi.com/10994300/21/5/479 |
| 45 | 2014 | Zixing Zhang; Eduardo  Coutinho; Jun Deng;  Björn Schulle | Active Learning  Semi-Supervised Learning  Cooperative Learning | Combining Active and SemiSupervised Learning Human-machine labeling collaboration based on confidence levels Test runs on emotion recognition tasks with varying initial training instances and feature sets | Cooperative Learning consistently outperforms individual techniques  Co-Training method achieves performance equivalent to full training set with 75% fewer labeled | Efficiently reduces need for human annotations  Robust method for leveraging unlabeled data in emotion recognition tasks. | 10.1109/TASLP.  2014.2375558 | https://ieeexplore.ieee.  org/abstract/document/697121 0/ |
| 46 | 2021 | Leonardo Pepino,  Pablo Riera, Luciana  Ferrer | Transfer learning with wav2vec 2.0 embeddings Simple neural networks for feature modeling Trainable weights for combining pre-trained model layers | Features from pre-trained models combined using trainable weights  Evaluation on IEMOCAP and  RAVDESS emotion databases  Comparison of wav2vec 2.0 models with and without finetuning for speech recognition | Superior performance compared to literature results Evaluation based on accuracy and other standard metrics | Transfer learning with wav2vec 2.0 embeddings enhances emotion recognition from speech. | 10.48550/arXiv.  2104.03502 | https://arxiv.org/abs/2104. 03502 |
| 47 | 2021 | Mustaqeem, Soonil Kwon | One-dimensional dilated  convolutional neural network  (DCNN)  Residual blocks with skip connection (RBSC) module Sequence learning (Seq\_L) module  Fusion layer for feature concatenation | Multi-learning strategy for spatial and contextual feature extraction  Fusion of learned features for emotion recognition  Lightweight architecture for real-time processing | High recognition accuracy:  73% on IEMOCAP, 90% on EMO-DB  Efficient real-time SER system implementation | Proposed MLT-SER system effectively learns emotional features from speech signals. Dynamic fusion framework enhances recognition rate by leveraging complementary features.  MLT approach reduces inefficiency and complexity of traditional CNN models. | 10.1016/j.eswa.  2020.114177 | https://www.sciencedirect.  com/science/article/pii/S09574  17420309131 |
| 48 | 2019 | Anjali Bhavan , Pankaj  Chauhan b, Hitkul ,  Rajiv Ratn Shah | Support Vector Machines  (SVMs) with Gaussian kernel Ensemble learning (bagging) | Feature extraction: spectral features  Dimensionality reduction:  Boruta  Feature selection: wrapperbased method | Bagged SVM ensemble outperformed boosting on emotion recognition from speech.  Reduction in feature dimensionality while maintaining recognition rate. Superiority demonstrated on  EmoDB, RAVDESS, and | Proposed bagged SVM ensemble method shows enhanced performance in speech emotion recognition. Acknowledges funding support from Infosys Center for AI and ECRA Grant. | 10.1016/j.knosys.  2019.104886 | https://www.sciencedirect.  com/science/article/pii/S09507  05119303533 |

IITKGP-SEHSC databases.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 49 | 2021 | Omar Mohamed, Salah  A. Aly | Wav2vec2.0 and HuBERT deep learning models  Multi-layer Perception (MLP)  Classifier  Bi-directional Long ShortTerm Memory (Bi-LSTM)  Layer | Feature extraction using  Wav2vec2.0 and HuBERT  Classification using MLP  Classifier and Bi-LSTM Layer Evaluation using F1 score, validation loss, and confusion matrix | Wav2vec2.0 outperforms HuBERT in accuracy and stability.  Wav2vec2.0 achieves 89% accuracy on Arabic BAVED dataset.  HuBERT base and large models achieve 87% and  84% accuracy, respectively. | Wav2vec2.0 outperforms HuBERT in accuracy and stability.  Wav2vec2.0 achieves 89% accuracy on Arabic BAVED dataset.  HuBERT base and large models achieve 87% and  84% accuracy, respectively. | 10.48550/arXiv.  2110.04425 | https://arxiv.org/abs/2110.  04425 |
| 50 | 2019 | Haiyang Xu, Hui  Zhang, Kun Han, Yun  Wang, Yiping Peng,  Xiangang Li | Attention mechanism  Sequential model for emotion recognition  IEMOCAP dataset | Learning alignment between speech frames and text  words using attention mechanism  Sequential model integration for emotion recognition from multimodal features | State-of-the-art performance achieved on the IEMOCAP dataset | Attention-based alignment improves multimodal emotion recognition from speech and text. | 10.48550/arXiv.  1909.05645 | https://arxiv.org/abs/1909. 05645 |
| 51 | 2020 | Siddique Latif, Rajib Rana, Shahzad  Younis, Junaid Qadir,  Julien Epps | Deep Belief Networks  (DBNs), Sparse  Autoencoder, Support Vector  Machine (SVM) | The paper utilizes transfer learning techniques to improve speech emotion recognition in cross-language and cross-corpus scenarios | DBNs achieved better accuracy than previous approaches, and using multiple languages for training with a fraction of target data significantly increased accuracy. | The study concludes that DBNs have superior feature learning abilities and can enhance baseline accuracy for practical applications, even with limited target data. Future work could focus on building robust systems for emotion recognition across multiple languages. | 10.48550/arXiv.  1801.06353 | https://arxiv.org/abs/1801. 06353 |
| 52 | 2021 | Author PictureSarala  Padi, Author  PictureSeyed Omid Sadjadi, Author  PictureRam D. Sriram,  Author PictureDinesh  Manocha | Transfer Learning and  Spectrogram Augmentation | TL and SA | Experimental results indicate that the transfer learning and spectrogram augmentation approaches improve the SER performance, and when combined achieve state-ofthe-art results. | In addition, we adopt a spectrogram augmentation technique to generate additional training data samples by applying random time-frequency masks to logmel spectrograms to mitigate overfitting and improve the generalization of emotion recognition models. | 10.1145/3462244.  3481003 | https://dl.acm.org/doi/abs/10. 1145/3462244.3481003 |
| 53 | 2018 | Albanie, S., Nagrani,  A., Vedaldi, A., & Zisserman, A. | The document discusses Cross-Modal Transfer in the context of Emotion  Recognition in Speech. | The document discusses Cross-Modal Transfer in the context of Emotion  Recognition in Speech. | accuracy: 72.6% | We have demonstrated the value of using a large dataset of emotion unlabelled video for cross-modal transfer of emotions from faces to speech. The benet is evident in the results – the speech emotion model learned in this manner achieves reasonable classication performance on | 10.1145/3240508.  3240578 | https://dl.acm.org/doi/abs/10. 1145/3240508.3240578 |
| 54 | 2017 | Aharon Satt, Shai  Rozenberg, Ron Hoory | Deep learning applied directly to spectrograms for  emotion recognition from speech | Deep neural network comprising convolutional and recurrent networks | Achieved a prediction accuracy of 66% over four emotions with a convolutiononly network, and 68% with a combined convolution-LSTM network | The proposed system demonstrates state-of-the-art accuracy in emotion recognition from speech with limited latency and robustness to background noise | 10.21437  /Interspeech.2017200 | https://www.isca-archive.  org/interspeech\_2017/satt17\_i nterspeech.pdf |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 55 | 2019 | Li, Y., Baidoo, C., Cai, T., & Kusi, G. A. | Speech Emotion Recognition | 1D CNN (One-Dimensional  Convolutional Neural  Network) | NACNN: 85.1% | This paper proposes a model based on 1D convolutional network for speech emotion recognition. Features were mainly extracted by complementary features, then based on 1D to capture emotional features. The model is evaluated on three databases, Emo-DB, | 10.1109  /ICSEC47112.  2019.8974716 | https://ieeexplore.ieee.  org/abstract/document/897471 6/ |
| 56 | 2016 | Priyanka A. Abhang, Bharti W. Gawali,  Suresh C. Mehrotra. | The book discusses EEG- and Speech-Based Emotion Recognition Methods, including EEG images and speech processing methods. | It mentions the use of electroencephalograms (EEGs) for detecting emotions and brain-computer interface (BCI) applications. | The book provides details on EEG-based emotion  recognition and classification. | It proposes a EEG/speech fusion method for accurate detection of emotional states in EEG recordings. | ISBN:  9780128045312,  0128045310 | https://books.google. com/books? hl=en&lr=&id=o4t4CgAAQBAJ &oi=fnd&pg=PP1&dq=speech+ emotion+recognition+&ots=pE cpYnpO4n&sig=poIp6xgiofSK4HVlUzcq1CbLZI |
| 57 | 2018 | Jian Huang, Ya Li,  Jianhua Tao, Zheng  Lian | Long Short-Term Memory Neural Network (LSTM),  Support Vector Machine  (SVM) | Triplet framework with LSTM for speech emotion recognition, trained with triplet loss and supervised loss12 | Improved performance on the IEMOCAP database; best accuracy achieved with cycle mode at 60.4% | The paper proposes a new approach to enhance speech emotion recognition and suggests further research for performance improvement | 10.21437  /Interspeech.20181432 | https://www.isca-archive.  org/interspeech\_2018/huang18 b\_interspeech.pdf |
| 58 | 2017 | Shih, P.-Y., Chen, C.P., & Wang, H.-M. | SER Using Skew-robust neural networks | Skew-robust neural networks | spkr-DNNs: 40.1% | We investigate a skew-robust parameter learning method for neural networks. The main idea is to introduce weights to training examples, which effectively modifies the traversed trajectory in the parameter space during the learning process. In addition, we apply cross-speaker | 10.1109/ICASSP.  2017.7952657 | https://ieeexplore.ieee.  org/abstract/document/795265 7/ |
| 59 | 2020 | Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Mansor, H., Kartiwi, M., & Ismail, N. | Speech Emotion Recognition using Convolution Neural Networks and Deep Stride  Convolutional Neural  Networks | SER Using CNN and DSCNNs | CNN: 79.7% accuracy DSCNN: 87.8% accuracy | The focus of Speech Emotion Recognition research is to design proficient and robust methods to recognize emotions. In this paper, we have modified the recently proposed algorithm Deep Stride Convolutional Neural Networks (DSCNN) by decreasing the number of | 10.1109  /ICWT50448.  2020.9243622 | https://ieeexplore.ieee.  org/abstract/document/924362 2/ |
| 60 | 2016 | Song, P., Ou, S., Zheng, W., Jin, Y., & Zhao, L. | SER | Transfer non-negative matrix factorization | The paper discusses the application of the technique. | In this paper, a new method called transfer non-negative matrix factorization is presented for cross-corpus speech emotion recognition. A graph based NMF approach is employed to  learn the robust representations of the  acoustic filter. | 10.1109/icassp.  2016.7472665 | https://ieeexplore.ieee.  org/abstract/document/747266 5/ |
| 61 | 2024 | Gaurav, Abhay Pratap Singh, Aadarsh Chaudhary | SVM,  ANN,  MLP,  Librosa Python Library,  RAVDESS dataset | Our approach combines individually trained Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Multilayer Perceptrons (MLPs) through ensemble stacking for improved prediction accuracy. | By combining the predictions from individually trained artificial neural networks (ANNs), support vector machines (SVMs), and multilayer perceptrons (MLPs), we achieved an impressive **81% accuracy** in classifying emotions on the RAVDESS dataset for 6 emotions | Our approach of combining predictions from separately trained artificial neural networks (ANNs), support vector machines (SVMs), and multilayer perceptrons (MLPs) yielded an impressive **81% accuracy** in classifying emotions on the RAVDESS dataset.  **Future Scope:** While achieving 81% accuracy is commendable, there are avenues for further improvement:   1. **Ensemble Techniques**: Investigate ensemble methods such as bagging, boosting, or stacking to enhance model performance. 2. **Feature Engineering**: Explore additional audio features or transform existing ones to capture more nuanced emotional cues. 3. **Fine-Tuning Hyperparameters**: Optimize hyperparameters for each model to squeeze out additional performance gains. 4. **Data Augmentation**: Generate synthetic samples to augment the dataset and improve model robustness. |  | https://github.com/abhay1704/SER |