International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7743
***Article Received**: 18 May 2023 **Revised:** 16 July 2023 **Accepted:** 06 August 2023*

Comprehensive Study of Automatic Speech Emotion Recognition Systems

Rupali Kawade[1], Sonal Jagtap[2]
[1]Department of E&TC Engineering
G H Raisoni College of Engineering and Management
Wagholi, Pune, India
rupali2118@gmail.com
[2]Smt. Kashibai Navale College Of Engineering
Department of E&TC Engineering
Vadgaon(Bk), Pune, India
sonalkjagtap@gmail.com

Abstract— Speech emotion recognition (SER) is the technology that recognizes psychological characteristics and feelings from the speech signals through techniques and methodologies. SER is challenging because of more considerable variations in different languages arousal and valence levels. Various technical developments in artificial intelligence and signal processing methods have encouraged and made it possible to interpret emotions.SER plays a vital role in remote communication. This paper offers a recent survey of SER using machine learning (ML) and deep learning (DL)-based techniques. It focuses on the various feature representation and classification techniques used for SER. Further, it describes details about databases and evaluation metrics used for speech emotion recognition.

Keywords- Affective Computing, Speech Emotion Recognition, Deep Learning, Machine Learning, Speech Recognition.

I. INTRODUCTION

Speech is the most natural and straightforward interaction that consists of information and emotion[1]. Emotion helps to understand the human being best. Speech is an essential medium in which expression communicates feelings and attitudes. The researcher's significant task is to find a speech signal's emotional content and classify the speech utterance's emotions. Over the last decade, SER has been regarded as an important research field. Automatic SER deals with the recognition of emotions from the speech signal [2][3]. SER is widely used in human-machine interaction [4], lie detection in psychiatric diagnosis [5], behavioral analysis in call center conversation [6], understanding the criminal behavior [7], robotics [8], etc.

The speech signal is often characterized by two main dimensions: valence and arousal [9][10]. Arousal is an autonic activation level created by events that lie in calmness (low) and excitation (high). Valence represents the pleasantness that lies in negative and positive. Valence and arousal level slightly dependent on several factors such as gender, region, language, and race[11][12][13].Fig. 1 shows the two-dimensional valence and arousal space.

Happy, delighted, and excitement lie inside the first quadrant of arousal-valence space with high arousal and positive valence. The second quadrant represents the tense, anger, and frustrating emotions. Quadrant three includes the emotions having low arousal and negative valences such as depression, boredom, and tiredness. The fourth quadrant of arousal valence space consists of emotions related to relaxation, calmness, and content.

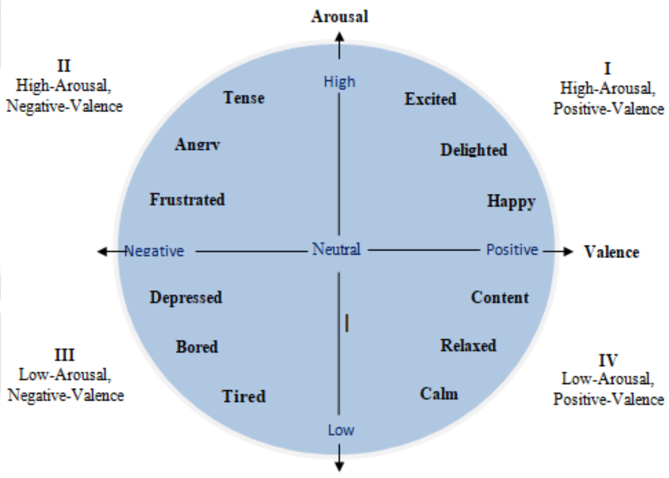

Fig. 1 Two dimensional valence-arousal space of emotion

The generalized framework of SER system encompasses speech pre-processing, feature representation, and classification stages, as shown in Fig. 2. The pre-processing stage deals with framing, normalization, cropping of speech signal, appending of speech signal, speech enhancement, noise removal, windowing, speech separation, voice activity detection, etc. [14]. Feature extraction captures the spectral ,prosodic, Teager energy operator (TEO) or voice quality features of the speech signal to improve the raw signal's discriminative power [15]. The speech signal can be classified into various emotions using ML or DL classifiers. The DL-based SER systems consider one-dimensional raw speech or two-dimensional representation of the raw speech signal as the input to deep network [16][17].

This paper offers a literature survey of recent machine learning and deep learning techniques for SER. It focuses on the various pre-processing methods, feature extraction algorithms, speech datasets, and performance assessment metrics. This paper summarizes the findings of recent literature on the SER and creates the background for the future improvement in the SER systems for potential researchers.

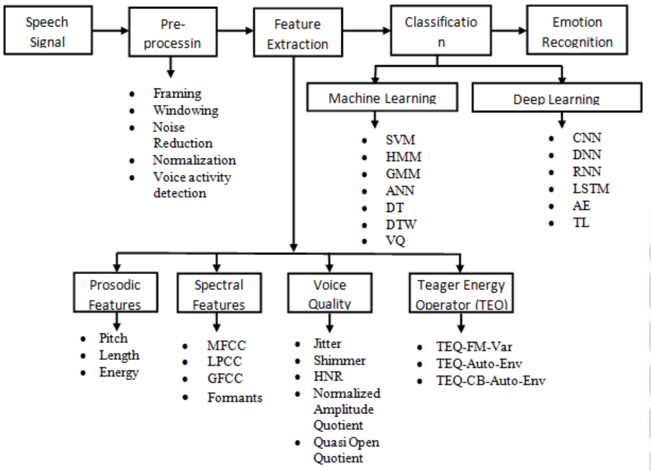

Fig. 2 Generalized framework of SER

The rest of the paper is structured as follow: Section II provides an ephemeral depiction of various feature extraction schemes utilized for SER. Section III provides a survey of SER based on machine learning. Next, section IV describes the recent survey of emotion recognition based on deep learning. Section V gives concise information related to different SER datasets. Further, section VI provides a detailed discussion about finding from a literature survey. Finally, section VII offers a conclusion and future direction for SER systems

II. FEATURE EXTRACTION TECHNIQUES

Speech emotion signal is continuous time-domain information and emotion. Speech characteristics might be local or global based on feature extraction. Local characteristics are segmental or short-term signal fluctuations. Long-term or supra-segmental global qualities reflect the signal's gross statistics. SER systems can assess local and global speech signals using spectral, prosodic, voice-quality, and Teager Energy Operator (TEO) aspects [18]. Intonation and rhythm affect prosody. Prosodic traits indicate happiness and anger more than fear and sadness. [19]. Alex et al. [20] trained deep neural networks using energy, duration, and fundamental frequency as speech and syllable prosodic characteristics. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database showed 63.83% un-weighted average recall (U.A.R.). Prosodic and spectral aspects enhance SER. [21]. Converting speech emotion data from time domain to frequency domain yields spectral properties. Vocal tract characteristics are spectral. Mel Frequency Cepstral Coefficients (MFCC) is a popular speech signal feature description method. It displays short-term voice signal power spectrum and phonemes as vocal tract shapes. Mel frequency scale matched perceived and real frequency [22][23] [24]. LPCC may approximate the vocal cord. MFCC outperforms LPCC in emotion recognition. LPC can efficiently encode low bit rate signals [25][26].Voice quality parameters often capture vocal tract physical attributes. Harmonics-to-noise ratio, shimmer, jitter, etc. Shimmer and Jitter reflect speech signal frequency and amplitude fluctuation. Jitter and shimmer assess frequency and amplitude instability [27].

TEO-based characteristics often recognise rage and stress. A non-linear vortex-airflow relationship in the vocal system shapes speech, according to Teager. Stress affects muscular rigidity and airflow during speaking [28][29]. TEO characteristics recognise tense emotions [30].

III. SER USING MACHINE LEARNING BASED TECHNIQUES

Support vector machine is popular for the SERbecause of its eager learning capability and lesser prediction time. Song et al. [31] presented SER. using Sparse Coding (S.C.) feature representation and transfer PCA-feature reduction. They have used Support Vector machine (SVM) classifier that resulted in accuracy of 43.97% (test-eNTERFACE, train-Berlin) and 51.36% (test-Berlin, train-eNTERFACE). It has shown better performance for the cross corpus SERbut shown less discrimination for unlabelled data. Further, Chen et al. [32] investigated technique that is appropriate to model human auditory perception based on Teager-Mel and P.L.P. features along with SVM classifier. It has shown 79.70% on the Chinese discrete emotional speech corpus (CDESC). Seho Lee [33] has presented combination of the naïve Bayes (NB) and K-Nearest Neighbor (KNN) to construct the NB-KNN that helps to minimize the over-fitting and attain fast convergence. It has given 69.20% and 68.4% accuracy for SAVEE and EMODB dataset respectively. Khan and Roy [34] presented SER based on MFCC and pitch features along with NB classifier. It has shown an accuracy of 81.00% accuracy but shows gender dependency. Sonawane et al. [35] suggested that non-linear SVM gives better performance compared to the linear SVM for the real time SER. The SVM classifier is widely utilized for the SER application because of its ability totransform the features in multi0dimensional plane to learn the distinctiveness of the raw features.

Feature extraction and feature selection are two main obstacles in cross-corpus SER. Feature Selection Based Transfer Subspace Learning (FSTSL) can acquiredistinctive low dimensional corpus independent features. They have extracted features using openSMILE toolkit and FSTSL [36].

Mao et al. [37] used LDA for redundancy minimization in INTER-SPEECH features for SER. It has shown 84.39% accuracy using SVM classifier on the Chinese speech database. Huang et al. [38] explored a Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient (LW-WPCC) to cope up with the issue of the additive noise in SER. It collaborates the short and long term frame analysis of the speech signal and provides noteworthy results when combined audio-visual representation for SER.

Dey et al. [39] have used LPC and LPCC for feature extraction and Golden Ratio based Equilibrium Optimization (G.R.E.O.) algorithm for feature selection. XGBoost classifier has given 97.31% and 98.46% on SAVEE and EmoDB dataset. This method has given better results compare with previous optimization techniques and deep learning techniques.

The inconsistency between training and testing samples decreases cross corpus emotion recognition rate. Sharing the source and target attribute space can lessen the incongruity between source and target features. Luo and Han [40] presented non-negative matrix factorization based transfer subspace learning method (NMFTSL) to improve the discrimination capability of features. It gave better performance for different language cross corpus SER. Sonmez et al. [41] applied Local Binary Pattern (LBP) and Local Ternary Pattern (LTP) for the SER. Initially 1-D Symlet wavelet filter is applied up to 9 levels to separate the high, medium and low level frequency components for speech. For each component LBP and LTP are applied. Further, features are fused together using min-max normalization algorithm and given to polynomial SVM classifier. It has low computational complexity but gives poor performance due to noisy data and class imbalance problem due to database inequalities.Kawade and Bhalke[42] presented wavelet packet coefficient (WPC) for the SER to acquire the diversity in the phonetic representation of the emotion signal. The WPC features and 1st and 2nd order differences of WPC shows robustness of the features for traditional ML based classifiers. The performance of WPC is evaluated for the KNN and SVM using correleation based feature selection technique on EMO-DB (80.85%) and RAVDESS (93.12%) dataset. Various ML techniques has been presented for the cross corpus SER which has shown promising results. The cross-corpus SER is still challenging due to difficulty in acquiring corpora of various languages, language intonation, regional and cultural effect on language. The pitch and energy of different language sentiments has shown wide variety which limits the effectiveness of cross corpus SER [43][44]. The success of traditional ML based SER highly rely on the feature extracted using feature extraction techniques. Its gives better performance for low sized database but gives deprived performance for larger sized database. Poor correlation and representation of the features degrades the SER accuracy.

IV. SER USING DEEP LEARNING BASED TECHNIQUES

In the recent years deep learning has attracted researcher's focus because of representation of highly discriminative features. Deep learning algorithms can be used for feature extraction as well as classification purpose. Zhang et al. [45] have presented Deep Convolutional Neural Networks (DCNN) to comprise gap between prejudiced emotions and low-level features for SER. The DCNN employed for learning MFCC features and SVM provided SER classification. The LP-norm pooling method outperforms average and maximum pooling. Neuman et al. [46] offered alternative CNN-based cross-lingual and multi-lingual SER (A.C.N.N.). Fine tuning with fewer parameters improves arousal prediction, whereas cross language training improves valence prediction. Multilingual training outperforms monolingual and cross-lingual training. Fine-tuning cross-lingual training may boost SER performance. Zhao et al. [47] used Merged DNN to analyse 1-D CNN and 2-D CNN speech spectrograms for SER. Bayesian optimization fine-tunes. It gave IEMOCAP 86.36% and EmoDB 89.77% accuracy. Ocquaye et al. [48] employed dual exclusive attentive transfer (DEAT) for unsupervised convolutional neural networks to adjust source and destination domains. CALLoss and second-order statistics of target and source attention mappings reduce domain discrepancy. 5-layered CNN uses voice spectrogram to enhance spectral characteristics and feature discrimination. . . Simple samples are taught initially in curriculum learning, followed by complicated samples. Fundamental frequency and MFCC extract features. It outperforms baseline approaches. Tripathi et al. [50] demonstrated SER using speech characteristics and transcripts. CNN learns emotion traits from MFCC and transcript text. It outperforms benchmarks by roughly 7%. Zhao et al. [51] stated that LSTM increases temporal feature representation and long-term feature reliance. 2D-CNN-LSTM outperforms 1D for subject-dependent and independent approaches. Peng et al. [52] learned emotion dynamics and cognitive continuity using 3D convolution and attention-based sliding recurrent neural networks (A.S.R.N.N.s). 3D convolution models periodic and local voice signals. A.S.R.N.N. represents local speech signal features. Attention outperformed maximum and mean pooling. Segment-based attention model outperformed frame-based. Data imbalance caused unsatisfactory results for MSP-IMPROV (Accuracy-55.70%) dataset. Conventional approaches cannot generalise and capture latent database information. Generalized domain adversarial neural network (GDANN) provides domain-invariant and generalised speech signal representation, and class-aligned GDANN (CGDANN) reduces class alignment issues caused by restricted labelled targets [53]. Redagging solves observation duplication and augagging full picture deficit. It reduced whole-picture and observation duplication issues. Xiaohan et al. [55] suggested DNN for temporal segment level low-level speech emotion signal characteristics. They employed energy, spectral, statistical, and voice-related low-level emotion signals. Aware temporal pooling outperformed average pooling. Chen et al. [56] introduced first-order attention network to address data imbalance and utterance variety. Pre-trained CNN (VGGish) network optimises log Mel spectrogram segment level characteristics. Bi-LSTM learns discriminative segment-level features. It reduced data imbalance and utterance variety. Smooth semi-supervised generative adversarial network (SSSGAN) was used to capture collaborative structure of labelled and unlabelled data for categorization. VSSSGAN reduced dependence on tagged data. It handles domain mismatch and data disturbances. Smoothing the adversarial model needed a bigger dataset [57]. The phase and loudness of speech reduce the frame clipping impact in SER. DRP and MCMA, based on deep CNN, provide promising results for SER on EmoDB and IEMOCAP datasets [58]. Bhangale et al. [59] described voice signal spectro-temporal features using 2-D Mel Frequency Logarithmic Spectrogram (MFLS). MFLS eliminates discrete cosine transform loss in the classic MFCC (DCT). Subject-dependent and subject-independent SER accuracy is 95.68% and 96.07% for the MFLS representation with lightweight DCNN. Kwon [60] suggested a lightweight 1-D dilated CNN-based multi learning technique (MLT) for voice signal spatio-temporal analysis. It helps SER blend salient and long-term contextual cues. Kwon [61] found that extensive feature selection improves SER accuracy. . 2-channel DCNN takes spectral and spatial features. INCA selects appropriate features and reduces severance by combining two-channel DCNN output. DL-based cross corpus SER approaches outperform ML-based ones because they correlate local and global speech signal characteristics better. Highly related features that bind the spectral, spatial, and temporal components of the emotion signal allow the DL algorithms transmit cross-language characteristics for cross-lingual SER [62-66]. Deep learning-based multiclass voice emotion recognition. It can accurately depict raw emotion signals. Deep learning outperforms machine learning. Deep learning methods are limited by architectural complexity, class-imbalance issue, longer training time, hyper-parameter tweaking complexity, etc.

IV. SPEECH EMOTION DATABASES

Emotion speech databases play an essential part in SER. Low quality, incomplete and faulty databases may degrade the SER. performance. Depending upon utterance styles speech emotion databases are categorized into acted (simulated), elicited (induced) and natural (spontaneous) speech databases. Utterances samples in an acted database are recorded by semi-professional or professional actors in a controlled environment. These samples are easy to record but often consist of exaggerated emotions that reduce the recognition rates for natural emotions. Elicited database samples are recorded by persons in a simulated emotional environment. Natural speech emotion database samples are collected from call center recording, talk show, radio show etc. Most of the databases are available in English and Mandarin languages only. Therefore, it is challenging to accommodate the existing systems for the new language.

V. EXPERIMENTAL RESULTS

We have implemented several machine learning based techniques for the SER on standard public EmoDB database. We evaluated performance of various classifiers such as Dynamic Time Warping (DTW), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), KNN, SVM, Naïve Bayes (NB), Linear Discriminant Analysis (LDA), Feed-forward Neural Network (FFNN) and Linear Vector Quantization (LVQ).We considered baseline spectral, time domain and voice quality features for the feature extraction without any pre-processing. As every ML classifier needs input features of same length, we cropped/ appended original EmoDB signals to 4sec. Fig. 3 shows the performance of SER for various ML based techniques for spectral features. The MFCC+SVM gives better performance for SER compared to other schemes without any pre-processing of data. Template based classifier like DTW and model based techniques such as HMM and GMM has given deprived results for SER. Eager learning techniques such as SVM, FFNN, NB, LDA, LVQ shows significant improvement in the results but its performance is limited because of lower feature correlation, lower inter-class variability and higher intra-class variability of the features.

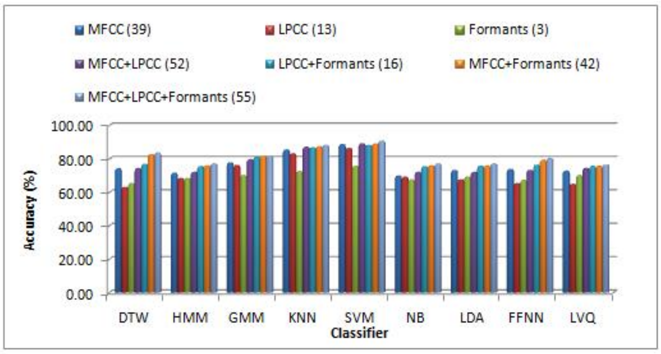

Fig. 3 Experimental results for spectral features for SER (EmoDB dataset)

It is noted that the MFCC based spectral features provides superior accuracy than LPCC and formants. However, combining different spectral features helps to improve the spectral domain representation of the signal and assists to achieve significant improvement in the SER accuracy. The ML classifiers performance is further validated for the time domain features such as zero crossing rate (ZCR) and pitch frequency; and voice quality features such as jitter and shimmer. It is observed that the voice quality features depicts better affective information provides better results compared with time domain features for EmoDB dataset as shown in Fig. 4. Further, the combination of different spectral, time domain and voice quality features are provided to ML classifiers for SER. The combination of different features provides improved feature representation due to emotional changes on the speech and helps to achieve significant improvement in accuracy as shown in Fig. 5.

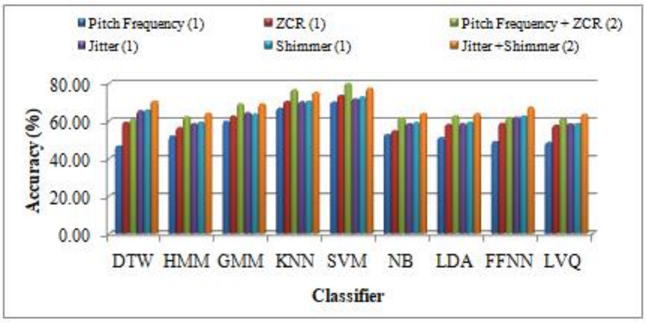

Fig. 4 Experimental results for time domain feature and voice quality features for SER (EmoDB dataset)

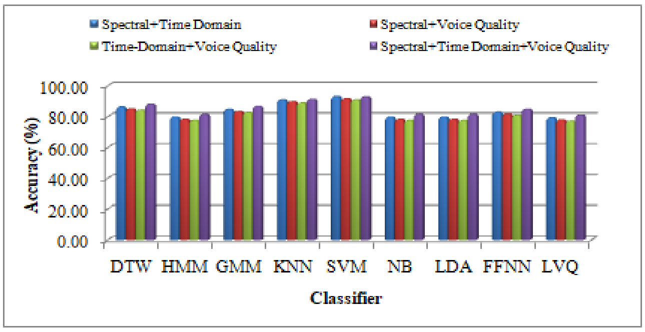

Fig. 5 Experimental results for time domain feature and voice quality features for SER (EmoDB dataset)

VI. CONCLUSION

This paper presents extensive survey of recent SER systems based on ML and DL techniques. The feature extraction is important stage for the SER based on machine learning techniques where performance of classified is greatly dependent on type, length and properties of the features. Deep learning techniques have improved the feature representation, correlation and internal temporal–spectral variations of the speech signal. Deep learning based emotion recognition systems have given better performance for noisy and larger database. Performance of SER system depends upon the size of database. Machine learning based techniques has given better performance for the small dataset whereas deep learning based techniques has provided better performance for the larger database. However, availability of large sized natural database and multilingual database is challenging. It is observed that cross-lingual and cross-corpus SER is still perplexing because of phonetic variations in the different languages. Most of the SER systems consider the speech phonetic parameters and neglects the actual context of the spoken content.

ACKNOWLEDGMENT

I would like to thank the Principal, Head of Department (E&TC) of Pimpri Chinchwad College of Engineering and Research, Ravet, Pune for their continuous support and motivation for the research.

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Communications of the A.C.M., vol. 61, no. 5, pp. 90–99, 2018.

[2] Bhangale, KishorBarasu, and K. Mohanaprasad. "A review on speech processing using machine learning paradigm." International Journal of Speech Technology 24, no. 2 (2021): 367-388.

[3] Bhangale, KishorBarasu, and MohanaprasadKothandaraman. "Survey of Deep Learning Paradigms for Speech Processing." Wireless Personal Communications (2022): 1-37.

[4] La Mura, Monica, and PatriziaLamberti. "Human-Machine Interaction Personalization: a Review on Gender and Emotion Recognition Through Speech Analysis." In 2020 IEEE International Workshop on Metrology for Industry 4.0 &IoT, pp. 319-323. IEEE, 2020.

[5] Bhangale, Kishor, and MohanaprasadKothandaraman. "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network." Electronics 12, no. 4 (2023): 839.

[6] Petrushin, Valery A. "Detecting emotion in voice signals in a call center." U.S. Patent 7,940,914, issued May 10, 2011.

[7] Ousmane, AbdoulMatine, TahirouDjara, and Antoine Vianou. "Automatic recognition system of emotions expressed through the face using machine learning: Application to police

interrogation simulation." In 2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART), pp. 1-4. IEEE, 2019.

[8] Chen, Luefeng, Wanjuan Su, Yu Feng, Min Wu, Jinhua She, and Kaoru Hirota. "Two-layer fuzzy multiple random forest for SERin human-robot interaction." Information Sciences 509 (2020): 150-163.

[9] Yu, Liang-Chih, Lung-Hao Lee, ShuaiHao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. "Building Chinese affective resources in valence-arousal dimensions." In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 540-545. 2016.

[10] Nicolaou, Mihalis A., HaticeGunes, and Maja Pantic. "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space." IEEE Transactions on Affective Computing 2, no. 2 (2011): 92-105.

[11] Li, Yongwei, Junfeng Li, and Masato Akagi. "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space." The Journal of the Acoustical Society of America 144, no. 2 (2018): 908-916.

[12] Parthasarathy, Srinivas, and Carlos Busso. "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes." arXiv preprint arXiv:1804.10816 (2018).

[13] Larradet, Fanny, RadoslawNiewiadomski, Giacinto Barresi, Darwin G. Caldwell, and Leonardo S. Mattos. "Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection." Frontiers in psychology 11 (2020): 1111.

[14] Akçay, Mehmet Berkehan, and Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, pre-processing methods, supporting modalities, and classifiers." Speech Communication 116 (2020): 56-76.

[15] Swain, Monorama, AurobindaRoutray, and PrithvirajKabisatpathy. "Databases, features and classifiers for speech emotion recognition: a review." International Journal of Speech Technology 21, no. 1 (2018): 93-120.

[16] Papakostas, Michalis, EvaggelosSpyrou, Theodoros Giannakopoulos, GiorgosSiantikos, DimitriosSgouropoulos, PhivosMylonas, and FilliaMakedon. "Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition." Computation 5, no. 2 (2017): 26.

[17] Özseven, Turgut. "A novel feature selection method for speech emotion recognition." Applied Acoustics 146 (2019): 320-326.

[18] Gao, Yuanbo, Baobin Li, Ning Wang, and Tingshao Zhu. "SERusing local and global features." In International Conference on Brain Informatics, pp. 3-13. Springer, Cham, 2017.

[19] L. Abdel-Hamid, N. H. Shaker and I. Emara, "Analysis of Linguistic and Prosodic Features of Bilingual Arabic–English Speakers for Speech Emotion Recognition," in IEEE Access, vol. 8, pp. 72957-72970, 2020, doi: 10.1109/ACCESS.2020.2987864.

[20] Alex, Starlet Ben, Leena Mary, and Ben P. Babu. "Attention and Feature Selection for Automatic SERUsing Utterance and Syllable-Level Prosodic Features." Circuits, Systems, and Signal Processing (2020): 1-29.

[21] Khan, Atreyee, and Uttam Kumar Roy. "Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier." In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), pp. 1017-1021. IEEE, 2017.

[22] Likitha, M. S., Sri Raksha R. Gupta, K. Hasitha, and A. Upendra Raju. "Speech based human emotion recognition using MFCC" In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), pp. 2257-2260. IEEE, 2017.

[23] Sonawane, Anagha, M. U. Inamdar, and Kishor B. Bhangale. "Sound based human emotion recognition using MFCC& multiple SVM." In 2017 International Conference on Information, Communication, Instrumentation and Control (I.C.I.C.I.C.), pp. 1-4. IEEE, 2017.

[24] Bhangale, Kishor B., Prashant Titare, RaosahebPawar, and SagarBhavsar. "Synthetic Speech Spoofing Detection Using MFCCAnd Radial Basis Function SVM." I.O.S.R. Journal of Engineering (I.O.S.R.J.E.N.), Vol. 8, Issue 6, pp.55- 62, 2018.

[25] Renjith, S., and K. G. Manju. "Speech based emotion recognition in Tamil and Telugu using LPCC.andhurst parameters—a comparitive study using KNN and ANN classifiers." In 2017 International conference on circuit, power and computing technologies (I.C.C.P.C.T.), pp. 1-6. IEEE, 2017.

[26] Feraru, Silvia Monica, and Marius Dan Zbancioc. "Emotion recognition in Romanian language using LPC features." In 2013 E-Health and Bioengineering Conference (E.H.B.), pp. 1-4. IEEE, 2013.

[27] Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, StefanosKollias, Winfried Fellenz, and John G. Taylor. "Emotion recognition in human-computer interaction." IEEE Signal processing magazine 18, no. 1 (2001): 32-80.

[28] Li, Xiang, and Xin Li. "SERUsing Novel HHT-TEO Based Features." J.C.P. 6, no. 5 (2011): 989-998.

[29] Drisya, P. S., and Rajeev Rajan. "Significance of teo slope feature in speech emotion recognition." In 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), pp. 438-441. IEEE, 2017.

[30] Bandela, Surekha Reddy, and T. Kishore Kumar. "Stressed SERusing feature fusion of teager energy operator and MFCC" In 2017 8th International Conference on Computing, Communication and Networking Technologies (I.C.C.C.N.T.), pp. 1-5. IEEE, 2017.

[31] Song, P., Zheng, W., Liu, J., Li, J., & Xinran, Z. (2015). A Novel SERMethod via Transfer P.C.A. and Sparse Coding. Chinese Conference on Biometric Recognition, 393-400.

[32] Chen, X., Li, H., Ma, L., Liu, X., & Chen, J. (2015). Teager Mel and P.L.P. Fusion Feature Based Speech Emotion Recognition. Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (I.M.C.C.C.), Qinhuangdao, 1109-1114.

[33] Lee, S. (2015). Hybrid Naïve Bayes K-nearest neighbor method implementation on speech emotion recognition. IEEE Advanced Information Technology, Electronic and Automation Control Conference (I.A.E.A.C.), Chongqing, 349-353.

[34] Khan, A. & Roy, U. K. (2017). Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier. International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 1017-1021.

[35] Sonawane, A., Inamdar, M. U. & Bhangale, K. B. (2017). Sound based human emotion recognition using MFCC& multiple SVM. International Conference on Information, Communication, Instrumentation and Control (I.C.I.C.I.C.), Indore, 1-4.

[36] Song, Peng, and Wenming Zheng. "Feature selection based transfer subspace learning for speech emotion recognition." IEEE Transactions on Affective Computing (2018).

[37] Mao, J., He, Y., & Liu, Z. (2018). SERBased on Linear Discriminant Analysis and Support Vector Machine Decision Tree. 37th Chinese Control Conference (CCC), Wuhan, 5529-5533.

[38] Huang, Y., Xiao, J., Tian, K., Wu, A., & Zhang, G. (2019). Research on Robustness of Emotion Recognition Under Environmental Noise Conditions. IEEE Access, 7, 142009-142021.

[39] Dey, Arijit, Soham Chattopadhyay, Pawan Kumar Singh, Ali Ahmadian, Massimiliano Ferrara, and Ram Sarkar. "A Hybrid Meta-Heuristic Feature Selection Method Using Golden Ratio and Equilibrium Optimization Algorithms for Speech Emotion Recognition." IEEE Access 8 (2020): 200953-200970.

[40] Dr. Govind Shah. (2017). An Efficient Traffic Control System and License Plate Detection Using Image Processing. International Journal of New Practices in Management and Engineering, 6(01), 20 - 25. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/52

[41] H. Luo and J. Han, "Nonnegative Matrix Factorization Based Transfer Subspace Learning for Cross-Corpus Speech Emotion Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2047-2060, 2020, doi: 10.1109/TASLP.2020.3006331.

[42] Sönmez, Yeşím Ülgen, and Asaf Varol. "A SERModel Based on Multi-Level Local Binary and Local Ternary Patterns." IEEE Access 8 (2020): 190784-190796.

[43] Kawade, Rupali, and D. G. Bhalke. "SERBased on Wavelet Packet Coefficients." In ICCCE 2021, pp. 823-828. Springer, Singapore, 2022.

[44] Zehra, Wisha, Abdul RehmanJaved, ZuneraJalil, Habib Ullah Khan, and Thippa Reddy Gadekallu. "Cross corpus multi-lingual SERusing ensemble learning." Complex & Intelligent Systems 7, no. 4 (2021): 1845-1854.

[45] Latif, Siddique, Adnan Qayyum, Muhammad Usman, and Junaid Qadir. "Cross lingual speech emotion recognition: Urdu vs. western languages." In 2018 International Conference on Frontiers of Information Technology (FIT), pp. 88-93. IEEE, 2018.

[46] Zhang, Shiqing, Shiliang Zhang, Tiejun Huang, and Wen Gao. "SERusing deep convolutional neural network and discriminant temporal pyramid matching." IEEE Transactions on Multimedia 20, no. 6 (2017): 1576-1590.

[47] Neumann, Michael. "Cross-lingual and multilingual SERonenglish and french." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (I.C.A.S.S.P.), pp. 5769-5773. IEEE, 2018.

[48] Zhao, Jianfeng, Xia Mao, and Lijiang Chen. "Learning deep features to recognise speech emotion using merged deep CNN." I.E.T. Signal Processing 12, no. 6 (2018): 713-721.

[49] Chaudhary, A. ., Sharma, A. ., & Gupta, N. . (2023). Designing A Secured Framework for the Steganography Process Using Blockchain and Machine Learning Technology. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 96–103. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2512

[50] Ocquaye, Elias NiiNoi, Qirong Mao, Heping Song, Guopeng Xu, and YanfeiXue. "Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition." IEEE Access 7 (2019): 93847-93857.

[51] Lotfian, Reza, and Carlos Busso. "Curriculum learning for SERfrom crowdsourced labels." IEEE/ACM Transactions on Audio, Speech, and Language Processing 27, no. 4 (2019): 815-826.

[52] Tripathi, Suraj, Abhay Kumar, Abhiram Ramesh, Chirag Singh, and PromodYenigalla. "Deep learning based emotion recognition system using speech features and transcriptions." arXiv preprint arXiv:1906.05681 (2019).

[53] Zhao, Jianfeng, Xia Mao, and Lijiang Chen. "SERusing deep 1D & 2D CNN LSTM networks." Biomedical Signal Processing and Control 47 (2019): 312-323.

[54] Peng, Zhichao, Xingfeng Li, Zhi Zhu, Masashi Unoki, Jianwu Dang, and Masato Akagi. "SERUsing 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends." IEEE Access 8 (2020): 16560-16572.

[55] Xiao, Yufeng, Huan Zhao, and Tingting Li. "Learning Class-Aligned and Generalized Domain-Invariant Representations for Speech Emotion Recognition." IEEE Transactions on Emerging Topics in Computational Intelligence (2020).

[56] Ai, Xusheng, Victor S. Sheng, Wei Fang, Charles X. Ling, and Chunhua Li. "Ensemble Learning With Attention-Integrated Convolutional Recurrent Neural Network for Imbalanced Speech Emotion Recognition." IEEE Access 8 (2020): 199909-199919.

[57] Xia, Xiaohan, Dongmei Jiang, and HichemSahli. "Learning Salient Segments for SERUsing Attentive Temporal Pooling." IEEE Access 8 (2020): 151740-151752.

[58] Chen, Gang, Shiqing Zhang, Xin Tao, and Xiaoming Zhao. "SERby Combining A Unified First-order Attention Network with Data Balance." IEEE Access (2020).

[59] H. Zhao, Y. Xiao and Z. Zhang, "Robust Semisupervised Generative Adversarial Networks for SERvia Distribution Smoothness," in IEEE Access, vol. 8, pp. 106889-106900, 2020, doi: 10.1109/ACCESS.2020.3000751.

[60] Guo, Lili, Longbiao Wang, Jianwu Dang, EngSiongChng, and Seiichi Nakagawa. "Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition." Speech Communication 136 (2022): 118-127.

[61] Bhangale, Kishor, and K. Mohanaprasad. "SERUsing Mel Frequency Log Spectrogram and Deep Convolutional Neural Network." In Futuristic Communication and Network Technologies, pp. 241-250. Springer, Singapore, 2022.

[62] Kwon, Soonil. "MLT-DNet: SERusing 1D dilated CNN based on multi-learning trick approach." Expert Systems with Applications 167 (2021a): 114177.

[63] Auma, G., Levi, S., Santos, M., Ji-hoon, P., & Tanaka, A. Predicting Stock Market Trends using Long Short-Term Memory Networks. Kuwait Journal of Machine Learning, 1(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/136

[64] Kwon, Soonil. "Optimal feature selection based SERusing two-stream deep convolutional neural network." International Journal of Intelligent Systems 36, no. 9 (2021b): 5116-5135.

[65] Neumann, Michael. "Cross-lingual and multilingual SERonenglish and french." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5769-5773. IEEE, 2018.

[66] Parry, Jack, Dimitri Palaz, Georgia Clarke, Pauline Lecomte, Rebecca Mead, Michael Berger, and Gregor Hofer. "Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition." In INTERSPEECH, pp. 1656-1660. 2019.

[67] Liu, Na, Yuan Zong, Baofeng Zhang, Li Liu, Jie Chen, Guoying Zhao, and Junchao Zhu. "Unsupervised cross-corpus SERusing domain-adaptive subspace learning." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5144-5148. IEEE, 2018.

[68] Su, Bo-Hao, and Chi-Chun Lee. "Unsupervised Cross-Corpus SERUsing a Multi-Source Cycle-GAN." IEEE Transactions on Affective Computing (2022).

[69] Braunschweiler, Norbert, Rama Doddipatla, Simon Keizer, and Svetlana Stoyanchev. "A study on cross-corpus SERand data augmentation." In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 24-30. IEEE, 2021.

[70] Burkhardt, F. ,Paeschke, A. , Rolfes, M. , Sendlmeier, W.F. , Weiss, B. , 2005. A database of German emotional speech.. In: Interspeech. ISCA, pp. 1517–1520

[71] Zhang, J.T.F.L.M. ,Jia, H. , 2008. Design of speech corpus for mandarin text to speech. The Blizzard Challenge 2008 workshop

[72] Kulkarni, L. . (2022). High Resolution Palmprint Recognition System Using Multiple Features. Research Journal of Computer Systems and Engineering, 3(1), 07–13. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/35

[73] Iemocap database. 2019. https://sail.usc.edu/iemocap/. Accessed: 2019-05-15.

[74] Surrey audio-visual expressed emotion database. 2019. https://sail.usc.edu/iemocap/ . Accessed: 2019-05-15.

[75] Toronto emotional speech database. 2019. https://tspace.library.utoronto.ca/handle/1807 /24487 . Accessed: 2019-05-15.

[76] Mao, X. , Chen, L. , Fu, L. , 2009. Multi-level SERbased on hmm and ann. In: 2009 W.R.I. World congress on computer science and information engineering, 7. IEEE, pp. 225–229 .

[77] Li, A. , Zheng, F. , Byrne, W. , Fung, P. , Kamm, T. , Liu, Y. , Song, Z. , Ruhi, U. , Venkatara- mani, V. , Chen, X. , 2000. Cass: A phonetically transcribed corpus of mandarin spontaneous speech. In: Sixth International Conference on Spoken Language Processing .

[78] Li, Y. , Tao, J. , Chao, L. , Bao, W. , Liu, Y. , 2017. Cheavd: a chinese natural emotional au- dio–visual database. J. Ambient Intell. Hum. Comput. 8 (6), 913–924 .

[79] Engberg, I.S. , Hansen, A.V. , Andersen, O. , Dalsgaard, P. , 1997. Design, recording and verification of a Danish emotional speech database. In: Fifth European Conference on Speech Communication and Technology .

[80] Ahammad, D. S. K. H. (2022). Microarray Cancer Classification with Stacked Classifier in Machine Learning Integrated Grid L1-Regulated Feature Selection. Machine Learning Applications in Engineering Education and Management, 2(1), 01–10. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/18

[81] Wang, K. , Zhang, Q. , Liao, S. , 2014. A database of elderly emotional speech. In: Proc. Int. Symp. Signal Process. Biomed. EngInformat., pp. 549–553 .

[82] Lee, S. ,Yildirim, S. , Kazemzadeh, A. , Narayanan, S. , 2005. An articulatory study of emo- tional speech production. In: Ninth European Conference on Speech Communication and Technology .

[83] Costantini, G. ,Iaderola, I. , Paoloni, A. , Todisco, M. , 2014. Emovo corpus: an italian emotional speech database. In: International Conference on Language Resources and Evaluation (L.R.E.C. 2014). European Language Resources Association (E.L.R.A.), pp. 3501–3504

[84] Gabriel Santos, Natural Language Processing for Text Classification in Legal Documents , Machine Learning Applications Conference Proceedings, Vol 2 2022.

[85] Martin, O. ,Kotsia, I. , Macq, B. , Pitas, I. , 2006. The enterface'05 audio-visual emo- tion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06). IEEE . 8–8.

[86] Mori, S. , Moriyama, T. , Ozawa, S. , 2006. Emotional speech synthesis using subspace con- straints in prosody. In: 2006 IEEE International Conference on Multimedia and Expo., pp. 1093–1096

[87] Liberman, M., Davis, K., Grossman, M., Martey, N., Bell, J., 2002. Emo- tional prosody speech and transcripts. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2002S28 Accessed: 2019-12-17.

[88] Ringeval, F. ,Sonderegger, A. , Sauer, J. , Lalanne, D. , 2013. Introducing the recola multi- modal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (F.G.). IEEE, pp. 1–8 .

[89] McKeown, G. ,Valstar, M. , Cowie, R. , Pantic, M. , Schroder, M. , 2011. The semaine database: annotated multimodal records of emotionally colored conversations be- tween a person and a limited agent. IEEE Trans. Affect. Comput. 3 (1), 5–17 .

[90] Hansen, J.H. ,Bou-Ghazale, S.E. , 1997. Getting started with susas: A speech under simu- lated and actual stress database. In: Fifth European Conference on Speech Communi- cation and technology

[91] Grimm, M. ,Kroschel, K. , Narayanan, S. , 2008. The vera am mittaggerman audio-visual emotional speech database. In: 2008 IEEE international conference on multimedia and expo. IEEE, pp. 865–868

[92] Batliner, A. ,Steidl, S. , Nöth, E. , 2008. Releasing a thoroughly annotated and processed spontaneous emotional database: the fauaibo emotion corpus. In: Proc. of a Satellite Workshop of L.R.E.C., 2008, p. 28

[93] Schuller, B. , Müller, R. , Eyben, F. , Gast, J. , Hörnler, B. , Wöllmer, M. , Rigoll, G. , Höthker, A. , Konosu, H. , 2009. Being bored? recognising natural interest by exten- sive audiovisual integration for real-life application. Image Vis. Comput. 27 (12), 1760–1774

[94] Kossaifi, J. ,Tzimiropoulos, G. , Todorovic, S. , Pantic, M. , 2017. Afew-va database for va- lence and arousal estimation in-the-wild. Image Vis. Comput. 65, 23–36 .

[95] Oflazoglu, C. ,Yildirim, S. , 2013. Recognizing emotion from turkish speech using acoustic features. E.U.R.A.S.I.P. J. Audio Speech Music Process. 2013 (1), 26

[96] Zhalehpour, S. ,Onder, O. , Akhtar, Z. , Erdem, C.E. , 2017. Baum-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans. Affect. Comput. 8 (3), 300–313