International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1S5, December 2019

Machine Learning Based Emotion Recognition using Speech Signal

k Ashok Kumar, J L Mazher Iqbal

Abstract: ***The challenging module in CAS (computer-aided services) has recognized the emotion from the signals of speech. In SER (speech emotion recognition), several schemes have used for extracting emotions from the signals, comprising various classification & speech analysis methods. This manuscript represents an outline of methods & explores some contemporary literature where the existing models have used for emotion recognition based on speech. This literature review presents contributions that made towards emotion recognition of speech and extracted the features for determining emotions.***

Keywords: ***Computer-Aided Services, Speech Emotion Recognition, Automatic Speech Recognition, Natural Language Processing, MFCC, SVM.***

I. INTRODUCTION

Regardless of the remarkable enhancement made in understanding natural language & speech, still we are not capable of communicating with the machines naturally. Designing a method for understanding human emotions could be dominant for several human interactions of computer applications. Nevertheless, it could be very challenging to design such methods. There are several modalities for expressing human emotions like body-posture, facial expression & voice. Hence, using manifold modalities might capture the expressed emotions accurately and result in optimal outcomes of recognition than unimodal methods [1]. Several contributions concentrated on utilizing modalities of audio-visual for recognition of emotion. This is because; both of them were very significant features of the expression of emotion. Nevertheless, in several real-time implementations, it was not possible for accessing data of audio-visual, and only data of audio is accessible. For instance, the recognition of emotion towards fatigue identification or call centers for drivers. Here, in such instance, the emotion recognition method utilizing speech signals could be promising. In every-day life, the sentence is uttered by a human in a general way that conveys the state of emotion through both contents & voice. Even though there were several contributions to the recognition of emotion in sentiment analysis & speech, only some of them researched collectively. Moreover, in conditions where only information of speech is available, one might use ASR (Automatic speech recognition) mechanism for converting signals of audio into text. Then, the multimodal method has applied for learning emotion from text & speech instantaneously.

Manuscript published on 30 December 2019.
* Correspondence Author (s)
K Ashok Kumar, Research Scholar, Veltech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai.
Dr. J. L. Mazher Iqbal, Professor, ECE, Veltech Rangarajan Dr. Sagunthala r&D, Institute of Science and Technology, Avadi, Chennai.

Hence, the text data were formed by the ASR mechanism that is trained generally from another huge dataset for speech recognition. Hence, it could argue that one can employ previous knowledge that has learned from other datasets aimed at recognizing the emotion. Here, this could deliberate as a transfer learning strategy, same as word embedding pre-training in NLP (Natural language processing) [2] or pre-training methods on imageNet for recognition of object [3]. For effectively utilizing both text data & speech, one requires to devise a method for collectively learning features from distinct fields. Even though some of the researches integrated both trainings & features a multimodal method, some of the contributions concentrated on the temporal association among text & speech at an optimal level. As text & speech are co-existed inherently in a temporal way, the multimodal method would be an advantage from utilizing alignment data. Usually, in the speech recognition method from the end-end, the method employs a technique that has a word to be decoded to get its resulting frames of speech [4], [5]. By this contribution, the focus is on learning the alignment between text & speech. Here, the text & speech features are integrated with word and serve to be multi-modal features aimed at an utterance of emotion. Even though an ASR model can result in alignment, our method does not need alignment from the ASR technique. This alignment could learn from attention techniques in this method. Here, there were 2 benefits for utilizing learned-alignment: primarily, our method is appropriate for an instance, where the ASR method is black-box & recognized text has resulted. For instance, utilizing google API speech recognition; second, learning the alignment for emotion recognition and might be optimal than alignment from the recognition of speech.

II. REVIEW OF CONTEMPORARY LITERATURE

This segment is mostly concerned about determining the emotion, representing its distinct methods. Various methods & inputs are utilized for emotions recognizing. An overview of the overall methods has depicted below.

1.1 Categorization of emotions

The emotion classification has been an interesting debate in distinct domains of affective-science, emotion research & psychology. Mostly, it is based on 2 well-known methods: dimensional & categorical. In the primary method, emotions were defined with a distinct amount of classes. The work [6] presents several theorists have carried out for defining that emotions were fundamental. In the 2nd method, [7] emotions were integration of various psychological dimensions & detected through axes. The other researchers determine emotions as per 1 or several dimensions.

Emotions could be defined by 3 dimensions: (a) relaxation versus strain, (b) unpleasurable& pleasurable & (c) subduing versus arousing [8]. The emotional state method PAD could be other 3-dimensional methods, where PAD denotes arousal, dominance & pleasure. The other well-known dimension method was proposed. The work [8] presents that unlike, former 3-dimensional methods, features of Russell's method only 2 dimensions that comprise (a) arousal & (b) valence [8]. The work [9] presents the categorical method that has generally utilized in SER. Here, it describes emotions utilized in daily emotion words like anger & joy. Here, in this contribution, the 6 fundamental emotions set plus neutral, resulting in 6 Ekman's method emotions, are utilized for emotion recognition from a speech by utilizing the categorical method. Voices were a significant modality for expressing the emotion. The speech could be an associated communication channel enhanced with the emotions: here, a voice in the speech would not only messages a semantic yet also gives information regarding speaker emotional state. Some significant vectors of voice feature, which have select for researching like MFCC (Mel-frequency cepstral coefficient).

1.2 Contemporary contributions

In the past decade, some of the researches utilize an integration of standard models. Two classification models have implemented for recognition of emotion of 6 emotional classes [10], where the performance classification is enhanced modestly when compared with the conventional model. The work [11] presents a 2-phase classifier for 5 emotions as projected, where SVM could utilize for classifying 5 emotions into 2 clusters. Here [12], HMMs were utilized for categorizing in binary-DT (decision tree) is utilized. The work [13] proposed a binary manifold classifier guided through dimensional emotion method. The work [14] [15] [16] presents that true comparison amid the outcomes of earlier mentioned models is very intricate because they might have utilized training, corpus and many more. Even though the models mentioned above are resourceful for detecting particular emotions, there could be no effective model for determining intricate emotional states, and the present speech recognition mechanism is immature still in humans' every-day life due to 2 reasons. Here, 1 reason could be speech emotion data redundancy that results to decrease in final recognition and training the sample data by taking a long time. Also, processing several speech data reduces the speed of the feedback system. The other reason could be that the effectiveness of the algorithm is based on independent features of the speaker that could implement for recognizing speech as it is not high. Besides, it also impacts the emotion recognition system of speech. In the above review, the emotion recognition of speech framework could project for lessening the impact of the individual variances and augment the feasibility of emotion recognition of speech. Primarily, the associative suitable emotional feature group of speech with independent characteristics of the speaker and emotional data could be chosen based on former researches. Then, the feature selection model is based on the Fisher criterion & correlation analysis has projected for lessening redundancy of feature. At last, the ELM-DT based on basic emotion confusion degree could represent the chosen representative emotional feature group of speech framework.

1.3 Speech Signal Features of Emotions

In a comprehensive sense as voice signals were not motionless, generally in handling the voice signal towards distinct voice motion into small divisions known as outlines. Internally in every casing, the flag could be viewed to be stationary roughly. The prosodic features of voice like vitality & pitch were disengaged from every edge and known as highlights. Later, the worldwide highlights were defined as overall highlights of voice signal measurements that have eradicated from expression. Here, there has been an inconsistency on which worldwide highlights & neighborhoods are appropriately increasing for acknowledgment of voice signaling. The work [17] presents that the vital part of specialists agreed that world-wide highlights were more optimal than a neighborhood in terms of grouping time & precision. The world-wide highlights have other ideal perspectives over neighborhood comprising; their amount is less deliberately. Likewise, the cross-approval using & highlight choice computations towards worldwide highlights were performed more rapidly than when linked with neighborhood-highlights. Nevertheless, the specialists have assured that the highlights of world-wide are effective in detecting excited feelings. For instance, dread, euphoria, outrage, trouble and many more [18]. They assure that the world-wide features disregard to cluster feelings that possess comparative excitement like happiness versus outrage. In worldwide features, the other drawback could be that fleeting data present in signal voice is gone completely. Moreover, it could be unreliable for utilizing conventional classifiers, for instance, HMM &SVM with signal features of worldwide as the preparing vectors quantity might not be sufficient for using an extensive amount of nearby comprised vectors and accordingly their specifications would evaluate precisely. The selected method needs to eradicate the element vector for every voice signal segment in spite of each phoneme. The segments of voice signal refer to voice constant signal parts that persuade by verbal line fluctuation and were oscillatory. Here, this technique is demanding less for actualizing that features those have based on the phoneme technique. The work [19] presents that a mixture of portion & worldwide features have accommodated by element vector. The algorithm of k-means clustering & SVM has used for analyzing.

1.4 Division of voice features

The important issue in acknowledgment of voice sentiments could be voice elicitation comprising that represents the enthusiastic voice substance, and in the meantime, it does not depend on lexical or speaker substance. Various signal highlights of voice have examined in the acknowledgment of the voice signal, yet specialists could not detect the optimal features of voice for the events. Here, the work [20] denotes features instances that have a place with each classification. Here, the main purpose of this domain is to examine the downsides & upsides of each classification.

1.5 Uninterrupted features of speech

Many of the researchers believe in prosody non-stop highlights, for instance, energy pass & pitch of required articulate substance. The work [21] presents that recently, some of the analysis affirmed towards the end. Determined signal highlights of voice have been used intensely in acknowledgment of voice signals. For a case, it could examine vocal signatures for 14 classifications of feeling [22]. The features of voice they used have detected by F0 (principal recurrence), explanation degree, vitality, & phantom information in voiced & unvoiced parts. The work [23] presents that analysis is done and acoustic highlights could assemble into accompanied classifications:

- The Features Associated to Pitch
- The Quality Features of Phonetic
- analogous features towards energy
- Features are Reckoned
- Features Pronunciation

Regularly, the world-wide features used in feeling acknowledgment of voice feelings are: mean, greatest, fluctuations, jitter, standard deviation, median, main distinction mean, and many more. In case of energy, median, least, min, max, standard deviation, average, & 4th request Legendre factors. In an instance of length, the ratio of voice & non-voice locales span, rate of voice signal & term of a wide signal of voice are the parameters. In the case of phonetics: 1st & 2nd phonetics and their speed of transfer. Likewise, the perplexing inputs are used extensively, for instance, in F0 design factors. Some of the analysis is done on the connection among earlier stated signal highlights of voice & required model feelings. For instance, bliss, dread, outrage, & astonishment feelings have attributes comparatively for F0 recurrence [24].

1.6 Features of voice characteristics

It is reliable that the enthusiastic articulation substance could detect emphatically with its characteristics of voice [25]. Exploratory analysis tunes into human subjects show a robust connection among apparent feeling & voice quality. The work [26] presents that several scientists deliberated the sound associated feelings parts that have been tried for connection characterize. The characteristics of voice through overall accounts represented in & out feelings. For instance, an individual feeling comes in the form of activities [23]. This could be in averse to fundamental feelings that affect adversely or emphatically individual activities without controlling clinching. An extensive opportunity of formants parameters adds to an abstract influence of characteristics of the voice. Here, acoustic relates, detected with the characteristics of voice are combined into classes.

- The dimension of voice: energy, term & adequacy of the flag have occurred as robust voice level proportions.
- Voice pitch
- The phoneme, feature limits, word & expression
- Temporal frameworks

To start with, the allusive marks were used for representing characteristics of voice like brutal, hoarse & rigid. The work [27] presents that those conditions might have distinct translations depend on scientist comprehension. This provoked a contradiction amid analysts for associating vocal characteristics circumstances for feeling. The work [26] presents that strained-voice could associate with happiness, fear & indignation & the voice remiss is associated with misery. Again, it has recommended that the voice of hoarse could associate with both satisfaction & displeasure; the trouble could associate with thunderous characteristics of the voice. The 2nd problem is an issue of selecting the conditions of voice characteristics consequently from the flag of voice. Several examinations are carried out on the last challenge that could arrange into 2 techniques. The significant technique depends on the way that the flag of voice could display in the form of vocal energized track channel by glottal source flag [24]. Consequently, the characteristics of voice might be predicted by abandoning the distinct vocal tract effect and predicting the glottal-flag factors.

In other instances, neither channel of vocal tract nor glottal source could depict, and hence, the glottal flag could be assessed by abusing data regarding source flag attributes and of the channel of the vocal tract. The work [27] presents the audit of the backward discerning process. In regard to the internal issues in this technique, it is not used extensively in acknowledgment of voice feeling. In the 2nd technique, the characteristics of voice are spoken empirically by factors specifically evaluated from voice-flag. For instance, no glottal flag source reckoning has executed. The work [28] presents that characteristics of sound would be spoken by shine & jitter. Some of the important features have utilized by deep learning methods at the time of several implementations like NLP, SER. In the instance of SER, many of these methods utilize supervised algorithms at the time of their application. Nevertheless, there could be a shift toward semi-supervised-learning [29]. Here, this might improve real-time data learning without the requirement of human manual labels.

The conventional SER methods usually include several classification methods like HMMs & GMMs. Here, GMMs are used to depict sound units of acoustic features. On other dimensions, HMMs are used to deal with the temporal changes in signals of speech [30] [31] [32]. Here, the modeling procedure utilizing such conventional methods needs a huge dataset for attaining accuracy in recognizing emotion. Therefore, it would be consuming more time. The work [33] presents that deep learning models are incorporated several non-linear rudiments, which execute computation in parallel [33]. Nevertheless, these models require to frame with deep layer structure.

The deep learning method is based on pre-training discriminative modality utilizing DNN-HMM besides with the coefficients of MFCC [34]. Here, DNN-HMM is been integrated with RBM by using unsupervised training for detecting distinct emotions of speech. The work [35] presents that hybrid modality of deep learning might attain better outcomes. A similar DNN-HMM could be depicted & compared to the GMM (Gaussian mixture method). Here, it examined along with RBM (restricted Boltzmann machine) in an instance where discriminative & unsupervised pre-training could be concerned more.

The outcomes attained in both instances have compared to those achieved from 2 layers & manifold layers of shallow NN-HMMs & GMM-HMMs. Here, DNN-HMMs that is hybrid with pre-training is having an accuracy of 12.22% by using a dataset called Enterface05 through an unsupervised training, 10,56% aimed at MLP-HMMs, 11.67% aimed at GMM-HMMs, & 17.22% aimed at shallow NN-HMMs in respective order. The work [36] presents that this recommends multimodality as an optimal avenue for researching, and also, there could be a time to enhance the emotion recognition accuracy, recognition system effectiveness & robustness [36].

The important issue, which impacts the overall RNN performance, could be sensitivity towards gradients disappearance [32] [33]. The work [34] presents an adaptive SER model based on a deep learning scheme called DRNN as utilized for the SER technique. Here, the learning phase of the method incorporates both short-time & frame-level acoustic features because of the same framework. The work [27] presents that other multitasking DNN through a shared hidden layer called MT-SHL-DNN have been used, where features transformation could be shared. The output layers are having a separate relationship with every utilized dataset. Here, DNN assists in scaling SER based on gender & speaker. If DNNs are utilized for segments encoding into vectors length, which are constant, then this could be done by utilizing several hidden layers pooling over a definite time. Here, the feature design encoding process could be done in a way that a classifier could jointly utilize it at a segmental level for effective classification. The work [37] presents that CNN (convolutional neural network) also utilizes a layer-wise framework and might classify the 7 universal-emotions from determined spectrograms of speech. The work [38] presents a method for SER, which is based on deep CNN & spectrograms. This method comprises of 3 fully integrated convolutional layers to extract emotional features from speech signal spectrogram images. The work [39] presents that other modifications, where method shares among target classes & associated sources have conducted. The 2-layered NN, where information of speech has gathered from several scenarios & sources are generally resulting in mismatching & therefore reduces the overall system performance. Primarily, the weights pre-training is conducted for 1st layer & then 2nd layer classification factors have enforced among 2 taken classes. Here, classes with minimum labeled data in the required domain might borrow data from related domain source for compensating defects.

The DNNs tendency is to learn particular features from several auditory emotion recognition methods [40]. Here, these features comprise recognition based on music & voice. Moreover, the cross-channel framework utilization might enhance the execution in an intricate environment. This method provides some optimal outcomes of music signal & speech signal of human; nevertheless, the outcomes for an auditory recognition of emotion is not optimum [41]. Here, the aid of this hierarchy of cross-section is to mine specific features & integrate them into more generalized cases. Besides, these methods could combine by DNNs based on visuals for enhancing the SER automatically. In such a case, the utilization of RNNs might further increase the execution of input information with dependent confines of time. The work [42] presents that CNN has evaluated by utilizing more autonomous cases called HRI (Human-robot interaction). Here, HRI has utilized as a robotic humanoid head, which resulted in required emotional-feedback.

The work [43], [44], [45] presents that the hybrid modality of deep learning might inherit the essential characteristics of RNN through CNN by the levels of convolutional implanted with the RNN. Here, this allows the method for attaining both temporal & frequency dependency in specified speech-signal. The work [45] presents that a part of memory improved RNN based on error reconstruction for incessant emotion recognition of speech. This method of RNN utilizes 2 components, primarily auto-encoder to rebuild features and, secondly, emotions estimating. It could utilize for attaining further inputs into RNN behavior based on BLSTM by utilizing regression methods like SVR [34].

The work [42] presents that SER algorithms have based on RNNs & CNNs. Here, a deep CNN hierarchical framework for the extraction of features has integrated with network layers of LSTM. It has identified that CNNs is having a distributed network based on time, which provides outcomes high maximum accuracy. The work [46] presents that the system is based on DCNN, which utilized input to be audio data called PCA-DCNNs-SER. Here, it comprised of 2 pooling & 2 convolutional-layers. The interferences in the background have been eradicated by utilizing the PCA (principal component analysis) strategy. The spectrogram of noise-free is segregated into components of non-overlapping. By utilizing acoustic features of hand-crafted, this method performs better classification based on SVM. Since expressions of emotion are spontaneous generally, their categorization into negative & positive domains would be simple.

The work [47] presents that preprocessing could be generally additional tasks that conduct before emotion identification from the signal of speech. For eradicating these confines, the SER method is required, which might contribute to the deep learning framework at an end-end. The work [35] presents that an instance is based on the processing spectrogram by utilizing DNN. It needs a deep hierarchical framework, augmentation & regularization of information.

The recognition of emotion from a speech by utilizing significant features is also contributed, where effective learning based on CNN is utilized [48]. Primarily, CNN would be trained by samples that are unlabeled for LIF (localized invariant features) learning using SAE (sparse auto-encoder). The important auto-encoder variation could be VAE (Variation auto-encoder) [47]. Later, this LIF could be used in the form of input for feature extraction by utilizing SDFA. The work [36] presents that the purpose of learning salient features, which are orthogonal & discriminative for emotion recognition of speech [36]. Here, the outcomes achieved with these simulations were more accurate, robust & stable in complex cases recognition where there could be a change in speaker & language, & other distortions in the environment. Emotions like joy, boredom, fear, sadness & anger could be identified. However, it becomes intricate to do when practical recognition of emotion is required. The change of this emotion learning based on salient features is depicted in [49], where semi –CNN would be represented for the SER.

Here, similar 2-phase training is conducted. Nevertheless, the simulation outcomes attained are stable, accurate & robust for recognition of emotion in cases like the distorted environment of speech. The simulations have conducted on artificial data. From the former decade, the deep learning scheme contributed prominent breakthroughs in understanding natural language.

The DBN (deep belief networks) for the SER [50], [51] exhibited a prominent enhancement over baseline methods [52] [53], which could not employ deep-learning. This recommends that non-linear associations of higher order are equipped better for recognition of emotion. The work [54] projected ELM (Extreme learning machine) that utilizes features at utterance level from probability distributions at segment level along with one hidden neural net layer for detecting emotions at utterance level, even though enhancement in accuracies were confined. The work [55] presents made utilization of hierarchical deep architectures, regularization & data augmentation with DNN for the SER, while the work [56] utilized spectrograms by Deep-CNNs. The work [57] trained DNNs on acoustic features sequence computed over little speech intervals besides with possible loss function of CTC that enabled long utterances deliberation comprising both unemotional & emotional parts and enhanced the accuracies of recognition. The work [49] utilized the bi-directional LSTM method for training the sequences of features and attained 62.8% of accuracy for emotion recognition on dataset IEMOCAP [58] that is having a prominent enhancement over the DNN-ELM [54]. The work [59] utilized deep-CNNs in integration with LSTMs for attaining better outcomes on data IEMOCAP.

In current years, the researchers are focusing on multimodal features utilization for recognition of emotion. The work [44] projected an SER model, which utilizes visual & auditory modalities for capturing emotional-content from numerous speaking styles. The work [60] projected a fusion tensor network that learns the dynamics of inter-modality & intra-modality end-end. The work [61] simulated with CDBN (convolutional deep belief networks) that learns salient expressions of multimodal features for attaining better accuracies. There is a huge literature for researching recognition of emotion by utilizing acoustic features [62]. The work [63] presents prosodic or spectral types LLDs (low-level descriptors) like pitch, intensity or MFBs (Mel-filter banks). The work [64], [65] presents that supra-segmental depictions have derived over language units like words, sentences, or phonemes proved as optimal than instantaneous segmental or frame-based methods. Typically, the features of supra-segmental have attained by computing several statistics from the LLDs over-determined language unit. Here, the set of feature sizes ranges often from 100 - 5000 relying on the amount of extracted statistics [66].

The discriminative classifiers were trained than over features at high-level for manifold tasks like categorical or binary classification & regression over incessant emotional features [67], [68], [69]. Contemporarily, the methods based on DNN have trained on speech features at frame-level that attained optimal performance than methods trained on utterance features a level of engineered humans [70], [71], [72]. Nevertheless, typically, these methods need training on huge tagged datasets.

On other dimensions, there were some of the researches that try for emotion characterizing by utilizing articulatory data. The work [73] exhibited that jaw degree opening enhanced in the form of subjects, whereas [74], the lateral lip distance among the mouth corners is exhibited to be impacted strongly by the emotional phase. The work [75] exhibits features based on articulation that attains the optimal rate of classification when compared with acoustic features for one male-subject. In every above contribution, the articulatory data has gathered by utilizing the manifold EMA (electromagnetic articulography) method.

These contributions are confined mostly towards one subject, or manifold speakers have recorded beneath the same circumstances. Nevertheless, these models need articulatory information at the time of inference. Such information acquisition on a huge scale could be intricate & consuming more time because of its highly sensitive & invasive recording process; this confines the application & scope of these models.

For alleviating the articulatory information requirement at the time of training, various researches projects to utilize pre-trained articulatory- acoustic inversion method for predicting articulatory features & later these features have utilized in distinct classification of tasks [76]. The work [77] presents that in an instance, the utility of the articulatory feature has predicted from an acoustic-articulatory pre-trained inversion method for the recognition of emotion. The work [78] presents that identical method has utilized for the phonetic-classification. Here, the work [79] combined the mapping system from acoustic-articulatory within the DNN framework and utilized this framework for both feature-mapping & senone classification. The work [80] presents that insight acoustic features have enhanced with predicted articulatory features produced from an acoustic-articulatory pre-trained method. The contribution projects that enhancing speech classification methods with predicted articulatory features outputs an enhanced performance. Nevertheless, these methods depend on intricate & non-linear methods of acoustic-articulatory, which are not better in the interest of the underlying task. The projected regularize in ACL could be suitable for databases by high-dimensional sets of features and confined articulatory information as same in the studies of emotion recognition. Here, in transfer learning, the aid is to enhance the required task performance by transmitting shared knowledge, which presents in associated source-task. Here, especially, this could be resourceful in implementations where training information is confined, but the dimension of feature is large [81]. Also, in this review, the transfer learning could be utilized typically in the form of adopting a method towards novel domains by comprising confined data from domain source at the time of training [82].

1.7 Results Description

This sction presents observations obtained from the contribution of this manuscript, which is the theortical analysis of the features speech signal towards emotion representation. The observations learned from the review carried on contemporary literature of emotion detection from speech signal can be listed as,

(i) Detection accuracy of the contemporary methods of emotion detection from speech signals are ascetically very low

(ii) considerable false alarming in emotion detection from speech signals.

(iii) The crux of the volume and dimensionality of the training corpus have not addressed

Hence, (i) the robust machine learning strategies for speech signal recognition, (ii) emotions centric features extraction and optimization, (iii) handling the crux of volume and dimensionality of the training corpus, and (iv) building comprehensive classifiers to perform supervised learning and predictive analytics towards empotion detection from speech signals are the significant objectives of the research.

III. CONCLUSION

This manuscript has delivered a detailed contemporary review of unsupervised & supervised learning methods for SER. Here, these contemporary models and their schemes are elaborated briefly based on distinct emotions classifications. These models deliver effective learning & predictive operations efficiently. The confines of these existing methods comprise learning from features at high dimensions and changing input data temporarily. This contribution works as fundamental for assessing the confines & performance of current unsupervised & supervised learning methods. Moreover, some favorable directions have highlighted for optimal SER approaches.

REFERENCES

1. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. InProceedings of the 28th international conference on machine learning (ICML-11) 2011 (pp. 689-696).
2. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems 2013 (pp. 3111-3119).
3. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015 Dec 1;115(3):211-52.
4. Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. InAdvances in neural information processing systems 2015 (pp. 577-585).
5. Chan W, Jaitly N, Le Q, Vinyals O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016 Mar 20 (pp. 4960-4964). IEEE.
6. Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Mahjoub MA. A review on speech emotion recognition: Case of pedagogical interaction in classroom. In2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) 2017 May 22 (pp. 1-7). IEEE.
7. Ekman P. An argument for basic emotions. Cognition & emotion. 1992 May 1;6(3-4):169-200.
8. Matilda S. Emotion recognition: A survey. International Journal of Advanced Computer Research. 2015;3(1):14-19
9. Koolagudi SG, Rao KS. Emotion recognition from speech: a review. International journal of speech technology. 2012 Jun 1;15(2):99-117.
10. Morrison D, Wang R, De Silva LC. Ensemble methods for spoken emotion recognition in call-centres. Speech communication. 2007 Feb 1;49(2):98-112.
11. Fu L, Mao X, Chen L. Speaker independent emotion recognition based on SVM/HMMs fusion system. In2008 International Conference on Audio, Language and Image Processing 2008 Jul 7 (pp. 61-65). IEEE.
12. Lee CC, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. Speech Communication. 2011 Nov 1;53(9-10):1162-71.
13. Xiao Z, Dellandréa E, Chen L, Dou W. Recognition of emotions in speech by a hierarchical approach. In2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops 2009 Sep 10 (pp. 1-8). IEEE.
14. Liu ZT, Li K, Li DY, Chen LF, Tan GZ. Emotional feature selection of speaker-independent speech based on correlation analysis and fisher. In2015 34th Chinese Control Conference (CCC) 2015 Jul 28 (pp. 3780-3784). IEEE.
15. Mao Q, Zhao X, Zhan Y. Extraction and analysis for non-personalized emotion features of speech. Advances in Information Sciences and Service Sciences. 2011;3(10).
16. Rybka J, Janicki A. Comparison of speaker dependent and speaker independent emotion recognition. International Journal of Applied Mathematics and Computer Science. 2013 Dec 1;23(4):797-808.
17. Slaney M, McRoberts G. BabyEars: A recognition system for affective vocalizations. Speech Communication. 2003 Feb 1;39(3-4):367-84.
18. Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. Speech communication. 2003 Nov 1;41(4):603-23.
19. Shami MT, Kamel MS. Segment-based approach to the recognition of emotions in speech. In2005 IEEE International Conference on Multimedia and Expo 2005 Jul 6 (pp. 4-pp). IEEE.
20. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 2011 Mar 1;44(3):572-87.
21. Johnstone T, Scherer KR. Vocal communication of emotion. Handbook of emotions. 2000 May 25;2:220-35.
22. Fernandez R. A computational model for the automatic recognition of affect in speech (Doctoral dissertation, Massachusetts Institute of Technology) 2004.
23. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG. Emotion recognition in human-computer interaction. IEEE Signal processing magazine. 2001 Jan;18(1):32-80.
24. L. Rabiner, R. Schafer, Digital Processing of Speech Signals, first ed., Pearson Education, 1978.
25. Scherer KR. Vocal affect expression: A review and a model for future research. Psychological bulletin. 1986 Mar;99(2):143.
26. Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America. 1993 Feb;93(2):1097-108.
27. Gobl C, Chasaide AN. The role of voice quality in communicating emotion, mood and attitude. Speech communication. 2003 Apr 1;40(1-2):189-212.
28. Li X, Tao J, Johnson MT, Soltis J, Savage A, Leong KM, Newman JD. Stress and emotion classification using jitter and shimmer features. In2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 2007 Apr 15 (Vol. 4, pp. IV-1081). IEEE..
29. GTrigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, Zafeiriou S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2016 Mar 20 (pp. 5200-5204). IEEE.
30. Huang KC, Kuo YH. A novel objective function to optimize neural networks for emotion recognition from speech patterns. In2010 Second World Congress on Nature and Biologically Inspired Computing (NaBIC) 2010 Dec 15 (pp. 413-417). IEEE.
31. Albornoz EM, Sánchez-Gutiérrez M, Martinez-Licona F, Rufiner HL, Goddard J. Spoken emotion recognition using deep learning. InIberoamerican Congress on Pattern Recognition 2014 Nov 2 (pp. 104-111). Springer, Cham.
32. Yu D, Seltzer ML, Li J, Huang JT, Seide F. Feature learning in deep neural networks-studies on speech recognition tasks. arXiv preprint arXiv:1301.3605. 2013 Jan 16.
33. Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Kingsbury B, Sainath T. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal processing magazine. 2012 Nov 1;29.
34. Tang D, Zeng J, Li M. An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals. InInterspeech 2018 (pp. 162-166).
35. Niu J, Qian Y, Yu K. Acoustic emotion recognition using deep neural network. InThe 9th international symposium on chinese spoken language processing 2014 Sep 12 (pp. 128-132). IEEE.

36. Li L, Zhao Y, Jiang D, Zhang Y, Wang F, Gonzalez I, Valentin E, Sahli H. Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In2013 Humaine Association Conference on Affective Computing and Intelligent Interaction 2013 Sep 2 (pp. 312-317). IEEE.
37. Zhang Y, Liu Y, Weninger F, Schuller B. Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations. In2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2017 Mar 5 (pp. 4990-4994). IEEE.
38. Badshah AM, Ahmad J, Rahim N, Baik SW. Speech emotion recognition from spectrograms with deep convolutional neural network. In2017 international conference on platform technology and service (PlatCon) 2017 Feb 13 (pp. 1-5). IEEE.
39. Mao Q, Xue W, Rao Q, Zhang F, Zhan Y. Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2016 Mar 20 (pp. 2608-2612). IEEE.
40. Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. In2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) 2016 Dec 13 (pp. 1-4). IEEE.
41. Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention. In2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017 Mar 5 (pp. 2227-2231). IEEE.
42. Barros P, Weber C, Wermter S. Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction. In2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids) 2015 Nov 3 (pp. 582-587). IEEE.
43. Lakomkin E, Zamani MA, Weber C, Magg S, Wermter S. On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks. In2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2018 Oct 1 (pp. 854-860). IEEE.
44. Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller BW, Zafeiriou S. End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing. 2017 Oct 18;11(8):1301-9.
45. Sahu S, Gupta R, Sivaraman G, AbdAlmageed W, Espy-Wilson C. Adversarial auto-encoders for speech based emotion recognition. arXiv preprint arXiv:1806.02146. 2018 Jun 6.
46. Zhang W, Zhao D, Chai Z, Yang LT, Liu X, Gong F, Yang S. Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services. Software: Practice and Experience. 2017 Aug;47(8):1127-38.
47. Mao Q, Dong M, Huang Z, Zhan Y. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE transactions on multimedia. 2014 Sep 29;16(8):2203-13.
48. Huang Z, Dong M, Mao Q, Zhan Y. Speech emotion recognition using CNN. InProceedings of the 22nd ACM international conference on Multimedia 2014 Nov 3 (pp. 801-804). ACM.
49. Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition. InSixteenth Annual Conference of the International Speech Communication Association 2015.
50. Kim Y, Lee H, Provost EM. Deep learning for robust feature generation in audiovisual emotion recognition. In2013 IEEE international conference on acoustics, speech and signal processing 2013 May 26 (pp. 3687-3691). IEEE.
51. Z Zheng WL, Zhu JY, Peng Y, Lu BL. EEG-based emotion classification using deep belief networks. In2014 IEEE International Conference on Multimedia and Expo (ICME) 2014 Jul 14 (pp. 1-6). IEEE.
52. Jin Q, Li C, Chen S, Wu H. Speech emotion recognition with acoustic and lexical features. In2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2015 Apr 19 (pp. 4749-4753). IEEE.
53. W Wu CH, Liang WB. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Transactions on Affective Computing. 2010 Dec 23;2(1):10-21.
54. Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine. InFifteenth annual conference of the international speech communication association 2014.
55. Fayek HM, Lech M, Cavedon L. Towards real-time speech emotion recognition using deep neural networks. In2015 9th international conference on signal processing and communication systems (ICSPCS) 2015 Dec 14 (pp. 1-5). IEEE.
56. Zheng WQ, Yu JS, Zou YX. An experimental study of speech emotion recognition based on deep convolutional neural networks. In2015 international conference on affective computing and intelligent interaction (ACII) 2015 Sep 21 (pp. 827-831). IEEE.
57. Chernykh V, Prikhodko P. Emotion recognition from speech with recurrent neural networks. arXiv preprint arXiv:1701.08071. 2017 Jan 27.
58. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation. 2008 Dec 1;42(4):335.
59. Satt A, Rozenberg S, Hoory R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. InINTERSPEECH 2017 (pp. 1089-1093).
60. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250. 2017 Jul 23.
61. Ranganathan H, Chakraborty S, Panchanathan S. Multimodal emotion recognition using deep learning architectures. In2016 IEEE Winter Conference on Applications of Computer Vision (WACV) 2016 Mar 7 (pp. 1-9). IEEE.
62. Shah M, Chakrabarti C, Spanias A. Within and cross-corpus speech emotion recognition using latent topic model-based features. EURASIP Journal on Audio, Speech, and Music Processing. 2015 Dec;2015(1):4.
63. Busso C, Lee S, Narayanan S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE transactions on audio, speech, and language processing. 2009 Mar 16;17(4):582-96.
64. Schuller B, Steidl S, Batliner A. The interspeech 2009 emotion challenge. InTenth Annual Conference of the International Speech Communication Association 2009.
65. Lee CC, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. Speech Communication. 2011 Nov 1;53(9-10):1162-71.
66. Eyben F, Wöllmer M, Schuller B. OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In2009 3rd international conference on affective computing and intelligent interaction and workshops 2009 Sep 10 (pp. 1-6). IEEE.
67. Parthasarathy S, Cowie R, Busso C. Using agreement on direction of change to build rank-based emotion classifiers. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016 Jul 22;24(11):2108-21.
68. Huang Z, Epps J. A PLLR and multi-stage staircase regression framework for speech-based emotion prediction. In2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2017 Mar 5 (pp. 5145-5149). IEEE.
69. Parthasarathy S, Busso C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. InINTERSPEECH 2017 Aug (pp. 1103-1107).
70. Le D, Aldeneh Z, Provost EM. Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network. InINTERSPEECH 2017 Aug (pp. 1108-1112).
71. Khorram S, Aldeneh Z, Dimitriadis D, McInnis M, Provost EM. Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition. arXiv preprint arXiv:1708.07050. 2017 Aug 23.
72. Aldeneh Z, Provost EM. Using regional saliency for speech emotion recognition. In2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017 Mar 5 (pp. 2741-2745). IEEE.
73. Erickson D, Fujimura O, Pardo B. Articulatory correlates of prosodic control: Emotion and emphasis. Language and Speech. 1998 Jul;41(3-4):399-417.
74. Nordstrand M, Svanfeldt G, Granström B, House D. Measurements of articulatory variation in expressive speech for a set of Swedish vowels. Speech Communication. 2004 Oct 1;44(1-4):187-96.
75. Lee S, Yildirim S, Kazemzadeh A, Narayanan S. An articulatory study of emotional speech production. InNinth European Conference on Speech Communication and Technology 2005.
76. Ghosh PK, Narayanan S. A generalized smoothness criterion for acoustic-to-articulatory inversion. The Journal of the Acoustical Society of America. 2010 Oct;128(4):2162-72.

77. J. Kim, P. Ghosh, S. Lee, S. S. Narayanan, in Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific. A study of emotional information present in articulatory movements estimated using acoustic-to-articulatory inversion (IEEE, 2012), pp. 1–4
78. Ghosh PK, Narayanan S. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. The Journal of the Acoustical Society of America. 2011 Oct 14;130(4):EL251-7.
79. Badino L, Canevari C, Fadiga L, Metta G. Integrating articulatory data in deep neural network-based acoustic modeling. Computer Speech & Language. 2016 Mar 1;36:173-95.
80. Li M, Kim J, Lammert A, Ghosh PK, Ramanarayanan V, Narayanan S. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Computer speech & language. 2016 Mar 1;36:196-211.
81. Tu M, Berisha V, Liss J. Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. InINTERSPEECH 2017 Aug (pp. 1849-1853).
82. Gideon J, Khorram S, Aldeneh Z, Dimitriadis D, Provost EM. Progressive neural networks for transfer learning in emotion recognition. arXiv preprint arXiv:1706.03256. 2017 Jun 10.