Grenze International Journal of Engineering and Technology, June Issue

A Survey on Deep Learning Methods for Speech Emotion Recognition

Lakshmi K.C[1], Dr.Sharath M N[2], Harshitha S[3] and Dr.Arjun B C[4]
[1]PG Student, Department of Computer Science and Engineering, Rajeev Institute of Technology, Hassan, India
Email: lakshmiqueen91@gmail.com
[2]Assistant Professor, Department of Computer Science and Engineering, Rajeev Institute of Technology, Hassan, India
Email: sharathmn64@rithassan.ac.in
[3]Assistant Professor, Department of Computer Science and Engineering, Malnad College of Engineering, Hassan, India
Email: sh@mcehassan.ac.in
[4]Professor & Head, Department of Information Science and Engineering, Rajeev Institute of Technology, Hassan, India
Email: bc.arjun@rithassan.ac.in

***Abstract*—Speech Emotion Recognition is a system that can recognize the emotion in various audio samples. Research in recent years has concentrated on applying deep learning to speech-related applications. This article provides an overview of deep learning techniques and examines work that uses these techniques to identify speech emotions. In this study, speech signals are used to analyze basic emotions such as calm, enjoyment, fear, disgust, etc. Review topics include the database employed, the emotions extracted, and the advancements made in speech emotion identification.**

***Index Terms*— ML Techniques, Speech Emotion Recognition, Deep learning, database.**

I. INTRODUCTION

Deep learning algorithms are typically employed to enhance computer capabilities so that they are capable of understanding what humans can accomplish, including speech recognition. Deep learning has been studied and used in a variety of different study subjects since it emerged as a new and appealing machine learning area in the previous ten years. [1] Emotions can be portrayed in speech, which can then be used to extract useful semantics from the uttered words, enhancing the effectiveness of speech recognition [2]. The strategy for speech emotion recognition (SER) is divided into two stages: feature extraction and feature classification [3].

The technique of turning raw data that hasn't been processed into numerical features that can be processed while preserving the specifics of the original data set is called feature extraction. The purpose of feature classification is to categorize each feature. Bayesian Networks (BN) or Support Vector Machines (SVM) are two of the most widely employed linear classifiers for speech emotion recognition. As an illustration, the Support Vector Machine (SVM) model is a classifier that uses mathematics to compute the audio signal's properties in order to predict the emotion. In the area of SER, this concept has had great success. However, the fundamental drawback of SVMs is that they can only divide the data into either class 1 or class 2 for classification. SVM drawbacks include lengthy training times and an inability to handle large datasets. Poor performance in areas with heavy noise and inability to identify local optimum.

A new instance is classified using the K-Nearest Neighbour (KNN) supervised learning algorithm based on the

Grenze ID: 01.GIJET.9.2.420

nearby training samples that are present in the feature space. The suggested KNN model divides the input signal into speech and music categories and most used in speech emotion recognition [5].

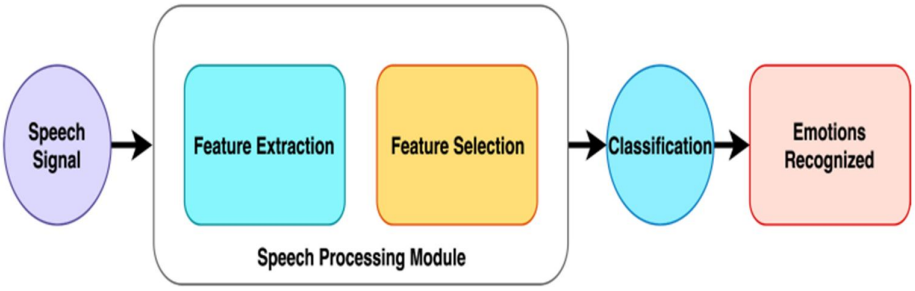

Figure 1: System for Detecting Emotions in Speech

The multi-layer perceptron model is regarded as the best model for identifying speech emotions. This is due to the fact that, when compared to other models like SVM, KNN, etc., the MLP classifier is more suitable for datasets with complicated structures [6]. In recent years, Deep Learning has drawn increased interest and is thought to be a developing field of research in ML [4]. On the other hand, speech-based classification techniques like natural language processing (NLP) and SER are greatly enhanced when recurrent structures like recurrent neural networks (RNNs) and long short-term memory (LSTM) [22].

II. Literature Review

In their review of speech recognition, J. Zhao et al. [7] used the Berlin EmoDb and IEMOCAP databases as well as deep learning techniques like CNN and DBN, with CNN LSTM achieving accuracy rates of 91.6% and 92.9%, respectively. S. Tripathi et al.'s model,[8] which was evaluated MoCap datasets and had an accuracy of 71.04%, was developed using IEMOCAP databases and deep learning techniques. C. W. Lee et al. [9] used the CMU-MOSEI dataset as well as deep learning techniques like LSTM-based CNN, and this model gives 83.11% accuracy in SER. For the purpose of identifying speech emotions, W. Zhang et al. [10] combined the CAS emotional speech database with deep learning techniques like SVM and Deep Belief Network (DBM). In comparison to SVM's accuracy of 84%, DBM provides a 94.61% accuracy rate.
Z. Lian Zhang et al. [11] used the CAS dataset with deep learning techniques like DBN and SVM, and the SVM model has an accuracy rate of 94.6%. Q. Mao et al. [12] used ABC, and Emo-DB databases as well as deep learning techniques like the Two-Layer Emotion-Discriminative and Domain-Invariant Feature Learning Method (EDFLM), and it gave the accuracy of 65.62% and 61.63%, respectively. J. Deng et. al [13] used ABC, EmoDb, and Geneva Whispered Emotion Corpus (GeWFC), as well as unsupervised universal autoencoders, form an adaptive model. This model learns discriminative data and incorporates the unlabelled learning with 62%, 63%, and 62.8% accuracy, respectively.
Z. Q. Wang and I. Tashev [14] used the Mandarin dataset as well as deep learning techniques like DNN-based extreme learning machines (ELM). This gives a weighted precision of 3.8% and 2.94% unweighted precision in speech emotion recognition. Pablo et al. [15] used the GTZAN, SAVEE, and EMotiW databases as well as the deep neural network (DNN). Jianwei et al. [16] used TIMIT corpus databases as well as DNN based acoustic features such as MFCCs, PLPs and FBANKs. 8.2% emotion recognition was obtained using 3 hidden layers based on DNN and acoustic data from MFCCs, PLPs, and FBANKs.
L. Li et al. [17] used the eNTERFACE'05 and Berlin databases as well as deep learning techniques like HMM-based Hybrid DNN (DNN-HMM), which provided 12.22% accuracy for the eNTERFACE'05 dataset. L. Fu et al [18] used Berlin database and mandarin database as well as deep learning techniques like Artificial Neural Network (ANN). Compared to absolute qualities, relative characteristics produced better results for emotion recognition.

III. Techniques for Speech Emotion Recognition (SER)

Fig 1 show SER system. The first stage of speech-based signal processing involves speech augmentation. The required features are extracted using a voice signal, and the choice is then made using the extracted features. At

the third stage, these features are categorised using a classifier, such as GMM and HMM. Last but not least, identifying emotions is dependent on feature classification.

TABLE I. VARIOUS EMOTION TYPES

Emotions	Pitch	Intensity	Speaking Time	Voice
Anger	unexpected amid stress	High	Faster	chest, breath
Fear	broad, typical	Low	Much faster	Unusual voicing
Happiness	an upward swivel	High	Faster/slower	Breath, shrill voice
Sadness	a tiny bit narrower	a downward swivel	Lower	Resonant

A. Improvement of SER Input Search Data

During the phase of data collection, a lot of noise taints the input data that if collected in order to identify emotions [19]. These flaws male feature extraction and categorization less precise [20]. During this pre-processing phase, the speaker and recording fluctuation are eliminated [21].

B. Extraction and Selection of Features

Considering the data gathered, pertinent traits are separated into numerous groups after extraction. Table 1 lists a few speech characteristics based on speech's acoustics and feelings.

C. About Features in SER

The two basic types of pattern recognition classifiers used for SER are often classified as linear classifiers and non-linear classifiers. Several conventional linear and non-linear classifiers are shown in table 2.

D. Databases used for SER

Simulated Databases. These databases contain data that has been captured by skilled and seasoned actors [29]. This strategy is thought to be used to compile more than 60% of all speech datasets. Another sort of database is an induced database, where the passionate set gathered by fabricating an emotional context [30]. Natural databases are the most realistic, but they are the hardest to gather because of the challenges associated with recognition [31]. Natural emotional speech dataset is typically compiled from recordings of discussions in the general public, contact centre interactions, and other settings.

TABLE II. FEW LINEAR AND NON-LINEAR CLASSIFIERS ARE EMPLOYED FOR SER

Classifier	Linear/Non-Linear(L/N-L)	References
SVM Classifier	L/N-L	[27],[28]
K-NN Classifier	L	[24]
HMM Classifier	N-L	[25],[26]

TABLE III. A COMPARISON OF SEVERAL CLASSIFIERS IN SER

Algorithms	Anger	Happy	Sad
K-NN	93%	55%	77%
SVM	74%	70%	93%
CNN	99%	99%	96%

E. SER Requires Deep Learning Techniques

Deep learning methods like the Deep Boltzmann Machine (DBM), the Recurrent Neural Network (RNN), the Recursive Neural Network (RNN), the Deep Belief Network (DBN), and the Convolutional Neural Network (CNN) methods used for SER, which significantly boosts the performance of the system that was designed. The main difference between the regular machine learning flow and the deep learning flow processes for SER is shown in Fig 2. Table 3 provides a comparison of Deep Learning, specifically the Deep Convolutional Neural Network (DCNN) algorithm [32].

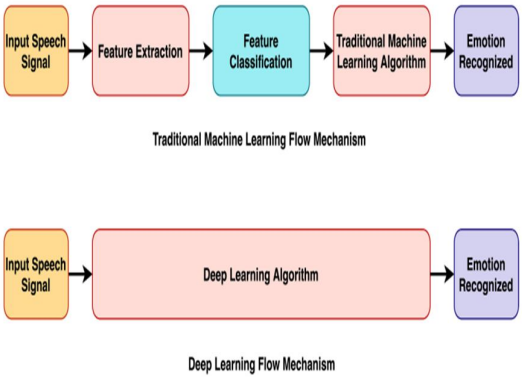

Figure 2 Traditional Machine Learning Flow vs Deep Learning Flow

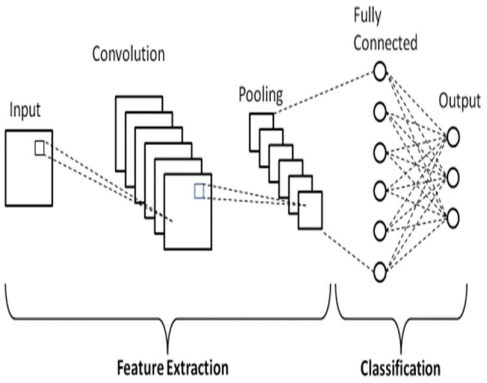

Figure 3 Convolutional Neural Network Architecture

IV. Methods of Deep learning for SER

A. Convolution Neural Network (CNN)

The most often used models now are CNN. One or more convolutional layers, which can either be completely connected or pooled, are present in this form of neural network computational model. Small size neurons that process the input data in the form of receptive fields are present on every layer of the intended model in these networks [33]. Fig 3 shows the architecture of CNN.

Advantages

- CNN offers high accuracy
- They are capable of automatically detecting features without any human control.

Disadvantages

- CNN does not encode an object's location or orientation.
- Large amount of training data is required

B. Recurrent Neural Network (RNN)

If the network forecast is inaccurate, each node in the RNN model functions as a memory cell, and the system learns and keeps working towards an accurate prediction during backpropagation. The RNN's sensitivity to gradient disappearance is the primary issue affecting its overall performance [34]. Fig 4 shows the architecture of RNN.

Advantages

- RNN keeps knowledge throughout time
- It is excellent for predicting time series called as LSTM (Long Short-term Memory)

Disadvantages

- Problems with gradient vanishing and exploding. RNN Training is extremely challenging task.
- If employing relu or tanh as an activation function, they are unable to handle very long sequences

C. Deep Belief Network (DBN)

A deep belief network is a form of deep neural network used in ML that consists layers of the latent variable and connections between the layers. DBNs are frequently used for SER due to their effectiveness in learning the

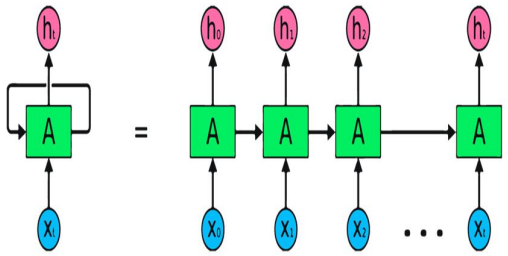

Figure 4 Architecture of Recurrent Neural Network

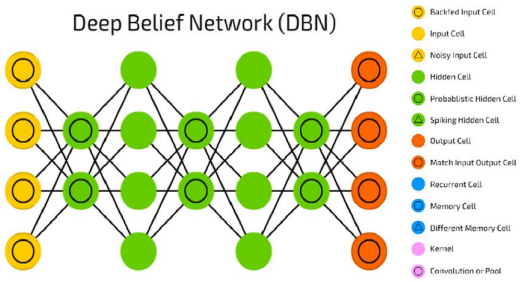

Figure 5 Deep belief network architecture

recognition parameters, regardless of how many parameters there are. Layer non-linearity is also avoided [35]. The layers then function as input feature detectors. DBNs can be thought of as a combination of straightforward, like autoencoders or limited Boltzmann machines. The hidden layer of each sub-network acts as the subsequent layer's visible layer. The DBN's architecture is depicted in Fig 5.

Advantages

- Effective at utilizing hidden layers.
- The same neural network methodology used in DBN can be applied to a variety of applications and data formats.

Disadvantages

- Requires a lot of data to operate.
- Due to its intricate data models, DBN is costly

V. Conclusions

The deep learning methods for SER have been thoroughly reviewed in this study. Deep learning methods like DBM, RNN, and CNN have attracted a lot of attention in recent years. Based on the classification of different natural emotions, these deep learning methods and their layer-wise designs are briefly described.

References

[1] H. Meftah, Y. A. Alotaibi, and S.-A. Selouani, ‘‘Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis,’’ IEEE Access, vol. 6, pp. 72845–72861, 2018.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, ‘‘Survey on speech emotion recognition: Features, classification schemes, and databases,’’ Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011

[3] S. G. Koolagudi and K. S. Rao, ‘‘Emotion recognition from speech: A review,’’ Int. J. speech Technol., vol. 15, no. 2, pp. 99–117, 2012.

[4] J. Schmidhuber, ‘‘Deep learning in neural networks: An overview,’’ Neu

[5] A. H. Meftah, M. Qamhan, Y. A. Alotaibi, and M. Zakariah, ‘‘Arabic speech emotion recognition using KNN and KSUEmotions corpus,’’ Int. J. Simul. Syst. Sci. Technol., vol. 21, no. 2, pp. 1–6, Mar. 2020.

[6] J. Joy, A. Kannan, S. Ram, and S. Rama, ‘‘Speech emotion recognition using neural network and MLP classifier,’’ IJESC, vol. 10, no. 4, pp. 25170–25172, Apr. 2020.

[7] J. Zhao, X. Mao, and L. Chen, ‘‘Speech emotion recognition using deep 1D & 2D CNN LSTM networks,’’ Biomed. Signal Process. Control, vol. 47, pp. 312–323, Jan. 2019.

[8] S. Tripathi and H. Beigi, ‘‘Multi-modal emotion recognition on IEMOCAP dataset using deep learning,’’ 2018, arXiv:1804.05788. [Online]. Available: https://arxiv.org/abs/1804.05788

[9] W. Lee, K. Y. Song, J. Jeong, and W. Y. Choi, ‘‘Convolutional attention networks for multimodal emotion recognition from speech and text data,’’ 2018, arXiv:1805.06606. [Online]. Available: https://arxiv.org/abs/1805.06606

[10] W. Zhang, D. Zhao, Z. Chai, L. T. Yang, X. Liu, F. Gong, and S. Yang, ‘‘Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services,’’ Softw., Pract. Exper., vol. 47, no. 8, pp. 1127–1138, 2017

[11] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, ‘‘Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN,’’ Sensors, vol. 17, no. 7, p. 1694, 2017

[12] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, ‘‘Learning emotiondiscriminative and domain-invariant features for domain adaptation in speech emotion recognition,’’ Speech Commun., vol. 93, pp. 1–10, Oct. 2017

[13] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, ‘‘Learning emotiondiscriminative and domain-invariant features for domain adaptation in speech emotion recognition,’’ Speech Commun., vol. 93, pp. 1–10, Oct. 2017
[14] Z.-Q. Wang and I. Tashev, ‘‘Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal (ICASSP), Mar. 2017, pp. 5150–5154.
[15] W. Lim, D. Jang, and T. Lee, ‘‘Speech emotion recognition using convolutional and recurrent neural networks,’’ in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA), Dec. 2016, pp. 1–4.
[16] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, ‘‘Deep neural networks for acoustic emotion recognition: Raising the benchmarks,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2011, pp. 5688–5691
[17] L. Fu, X. Mao, and L. Chen, ‘‘Relative speech emotion recognition based artificial neural network,’’ in Proc. IEEE Pacific–Asia Workshop Comput. Intell. Ind. Appl. (PACIIA), vol. 2, Dec. 2008, pp. 140–144.
[18] N. Morgan, ‘‘Deep and wide: Multiple layers in automatic speech recognition,’’ IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 7–13, Jan. 2012.
[19] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, ‘‘Reconstruction-errorbased learning for continuous emotion recognition in speech,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2367–2371.
[20] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, ‘‘A survey of affect recognition methods: Audio, visual, and spontaneous expressions,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009.
[21] T. Vogt, E. André, and J. Wagner, ‘‘Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation,’’ in Affect and Emotion in Human-Computer Interaction. Springer, 2008, pp. 75–91.
[22] J. Schmidhuber, ‘‘Deep learning in neural networks: An overview,’’ Neural Netw., vol. 61, pp. 85–117, Jan. 2015.
[23] S. Demircan and H. Kahramanlı, ‘‘Feature extraction from speech data for emotion recognition,’’ J. Adv. Comput. Netw., vol. 2, no. 1, pp. 28–30, 2014.
[24] O. Pierre-Yves, ‘‘The production and recognition of emotions in speech: Features and algorithms,’’ Int. J. Hum.-Comput. Stud., vol. 59, nos. 1–2, pp. 157–183, 2003.
[25] A. D. Dileep and C. C. Sekhar, ‘‘HMM based intermediate matching kernel for classification of sequential patterns of speech using support vector machines,’’ IEEE Trans. Audio, Speech, Language Process., vol. 21, no. 12, pp. 2570–2582, Dec. 2013.
[26] G. Vyas, M. K. Dutta, K. Riha, and J. Prinosil, ‘‘An automatic emotion recognizer using MFCCs and hidden Markov models,’’ in Proc. IEEE 7th Int. Congr. Ultra Mod. Telecommun. Control Syst. Workshops (ICUMT), Oct. 2015, pp. 320–324.
[27] Y. Pan, P. Shen, and L. Shen, ‘‘Speech emotion recognition using support vector machine,’’ Int. J. Smart Home, vol. 6, no. 2, pp. 101–108, 2012.
[28] Y. Chavhan, M. Dhore, and P. Yesaware, ‘‘Speech emotion recognition using support vector machine,’’ Int. J. Comput. Appl., vol. 1, no. 20, pp. 6–9, 2010.
[29] M. Swain, A. Routray, and P. Kabisatpathy, ‘‘Databases, features and classifiers for speech emotion recognition: A review,’’ Int. J. Speech Technol., vol. 21, no. 1, pp. 93–120, 2018.
[30] P. Jackson and S. Haq, Surrey Audio-Visual Expressed Emotion (SAVEE) Database. Guildford, U.K.: Univ. Surrey, 2014.
[31] R. Cowie, E. Douglas-Cowie, and C. Cox, ‘‘Beyond emotion archetypes: Databases for emotion modelling using neural networks,’’ Neural Netw., vol. 18, no. 4, pp. 371–388, 2005.
[32] M. Sidorov, S. Ultes, and A. Schmitt, ‘‘Emotions are a personal thing: Towards speaker-adaptive emotion recognition,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2014, pp. 4803–4807.
[33] Y. Kim, ‘‘Convolutional neural networks for sentence classification,’’ 2014, arXiv:1408.5882. [Online]. Available: https://arxiv.org/ abs/1408.5882
[34] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, ‘‘Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio,’’ in Proc. IJCAI, 2016, pp. 2196–2202.
[35] A. Mohamed, G. E. Dahl, and G. Hinton, ‘‘Acoustic modeling using deep belief networks,’’ IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 14–22, Jan. 2012.