

# **An Overview of Speech Emotion Recognition using Machine Learning Techniques**

Sushma Bahuguna  
Sr. Asst. Professor, CSE,  
BCIIT, Affiliated to GGSIPU, Delhi, India

## **I. INTRODUCTION**

Automatic Speech Recognition (ASR) System, infers human language, derives meaning from it, and performs the task for the user. Research and development in speech technology began to expand during the 2010's and has gained significant popularity for providing effective interaction between human and computers [1]. Microsoft's Cortana, Google's Google Assistant, and Apple's Siri, Amazon's Alexa, are a few examples of ASR system that employ speech recognition to Human-Computer Interaction. Though the ASR system is a powerful product of Human-Computer Interaction, the Speech Emotion Recognition (SER) is still a challenge as emotions are subjective and people might interpret it differently as there is no common consent on how to measure them [2]. Besides, there is no standard set of labels for human emotions. Also, there are many factors that may influence the SER like background noises, voice expressions, volume, cold, misspoken or misread prompted phrases, previous user activity etc. A classic SER system can be considered as a collection of procedures that isolate, extract and classify speech signals to detect emotions embedded in them [3]. Emotion plays a vital role in communication and its recognition and analysis can provide wide variety of applications to various institutions and aspect of life. Currently application of machine learning and neural network tool have got increasing attention in the research field of SER. This chapter gives a brief overview of different machine learning (ML) techniques that have been used for detecting emotional states in vocal expressions.

## **II. SPEECH EMOTION RECOGNITION**

The approach for Speech Emotion Recognition (SER) primarily comprises two phases known as feature extraction and features classification phase [4]. Figure 1 illustrates the phases of Speech Emotion Recognition system. The first phase relates to extraction of suitable features from speech data that are capable to characterize different emotions. Researchers have implemented several features such as, source-based excitation features, prosodic features, vocal traction factors, and other hybrid features in the field of speech processing technology [5]. Feature extraction is accomplished by changing the speech waveform to a form of parametric representation at a relatively minimized data for subsequent processing and analysis [6]. The second phase includes feature classification in which classifier decides the underlying emotion of the speech utterance. This phase characterizes the interface between the problem domain and the classification techniques. In literature different type of ML classifier such as single classifiers, multiple classifiers, and ensemble classifiers are implemented to develop SER system. Bayesian Networks (BN), Gaussian Mixtures Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random forest

(RF), Artificial Neural Network (ANN), etc., are some of the most commonly used classifiers in ERS [7-9].

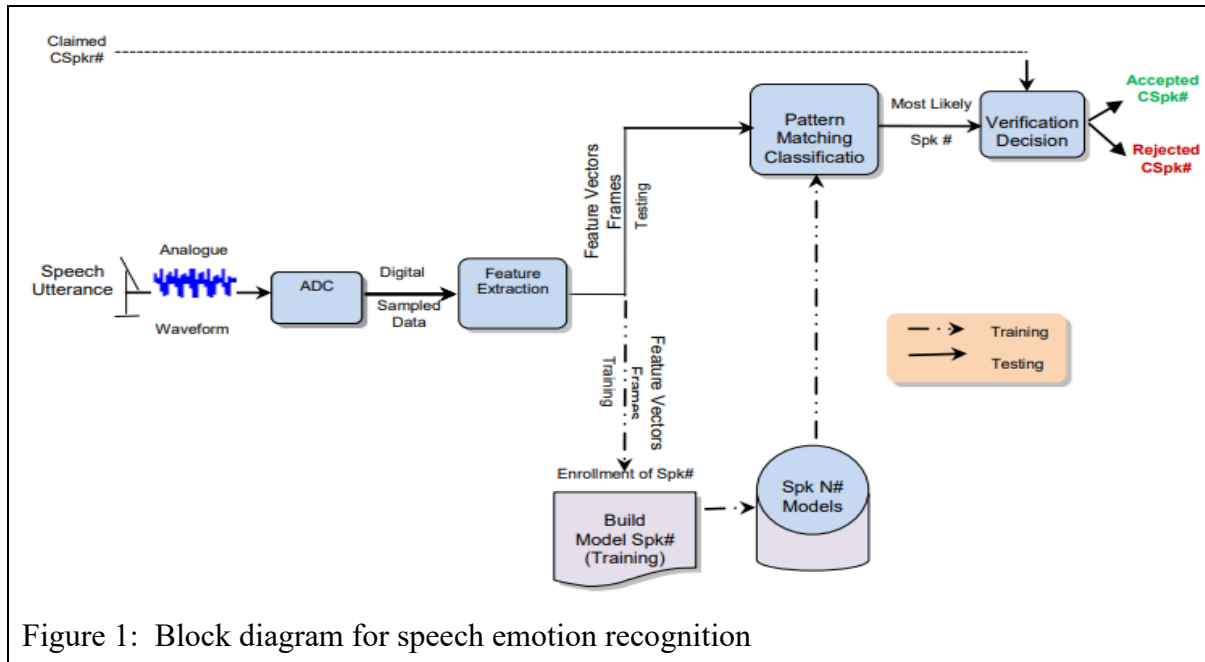


Figure 1: Block diagram for speech emotion recognition

### A. Feature Extraction Techniques

Feature extraction and production of parameters from the speech signal is one of the key steps in speech processing [10]. Feature extraction is implemented to emphasis on information in the signal, comparison between different classes and reduce the dimension of data and calculations [11]. Two classes of features namely prosodic features and characteristics of the vocal tract system have been used in SER. The first class is extracted from prosodic data such as Pitch, Energy, and Duration etc. The second class is allied to the vocal tract that includes Cepstral coefficients such as MFCC, LPCC, Formants etc. Spectral features extracted from vocal track such as Linear Predictive Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), etc. are most commonly used features for SER studies.

#### 1. Linear Predictor Coefficients (LPC)

LPC technology was advanced by Bishnu Atal and Manfred Schroeder during the 1970s–1980s [12]. LPC represents the characteristics of particular channel of speech. Different emotional speech of a person have different channel characteristics and coefficient of these features can be extracted for emotion recognition in the speech.

#### 2. Mel Frequency Cepstrum Coefficients (MFCC)

MFCC introduced by Davis and Mermelstein in the 1980's, have been state-of-the-art ever since [13]. MFCC uses a nonlinear frequency unit to simulate the human auditory system based on the characteristics of the human ear's hearing with good ability of the distinction, anti-noise and other advantages [14].

#### 3. Mel Energy-spectrum Dynamic Coefficients (MEDC)

MEDC extraction process is similar with MFCC except it takes logarithmic mean of energies after Mel Filter bank and Frequency wrapping instead of the logarithmic after Mel Filter bank and Frequency wrapping.

#### **4. Perceptual Linear Prediction (PLP)**

PLP was first proposed by Hynek Hermansky as a way of warping spectra to minimize the differences between speakers while retaining the important speech information [15]. The Differences between PLP and MFCC process lie in the filter-banks, the equal-loudness pre-emphasis, the intensity-to-loudness conversion and the application of LP, each of which makes PLP more consistent with human auditory impression [16].

#### **B. Feature Classifiers**

After extracting relevant features from speech data, classifier selects the underlying emotion of speech utterance [17]. In order to achieve the accurate emotional output, selection of suitable classifier is of vital importance. There is no specific criteria to select appropriate classifier and selection of classifier depends on the judgment of the researcher on the basis of geometry of the input vectors. Different classifier have been applied for SER and each classifier has its advantages and limitations. Following sections describe different classifier to SER, reported in the literature based on machine learning and neural network approaches.

##### **(i) Conventional machine learning approaches**

#### **1. Bayesian Networks (BN)**

The term Bayesian network was invented by Judea Pearl [18]. In the late 1980s Pearl's Probabilistic Reasoning in Intelligent Systems [19] and Neapolitan's Probabilistic Reasoning in Expert Systems [20] recognized Bayesian network as a field of study. BN is a probabilistic graphical model that represents a set of variables and their conditional dependencies using a directed acyclic graph. It consists of two parts namely Directed Acyclic Graph and Table of conditional probabilities. A BN network graph consists of nodes and Arcs. Each node represents random variables (continuous or discrete) and Arc represent the conditional probabilities or causal relationship between random variables. Each node in the Bayesian network has condition probability distribution that determines the effect of the parent on that node.

#### **2. Support Vector Machine (SVM)**

The SVM algorithm was developed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1964 and soft margin incarnation was proposed by Corinna Cortes and Vapnik in 1993 [21]. SVM is a supervised ML algorithm used for both classification and regression. The main aim of the SVM algorithm is to find the optimal hyper plane in an N-dimensional space that can separate the data points in different classes in the feature space. SVMs are flexible and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships.

#### **3. Gaussian Mixture Model (GMM)**

German mathematician Carl Friedrich Gauss introduced Gaussian distribution and made notable contributions to the field, gave his name to the model. GMM is a probabilistic model that assumes all the data points are produced from a mixture of a finite number of Gaussian distributions with unknown parameters. In general, models don't require knowing which subpopulation a data point belongs to, allowing the model to learn the subpopulations automatically. A Gaussian mixture model can be used for clustering, which is the task of grouping a set of data points into clusters.

#### **4. Hidden Markov Model (HMM)**

HMM was introduced by L.E. Baum in the late 1960s [22] and are popular in speech recognition research to model the statistical variation of the spectral features. HMM is a statistical model describes the probabilistic relationship between a sequence of observations and a sequence of hidden states. It is often used in situations where the underlying process that generates the observations is unknown. It predicts future observations or classify sequences on the basis of underlying hidden process that generates the data. It consists of two types of variables: hidden states (underlying variables that generate the observed data, but are not directly observable) and observations, are variables that can be measured and observed.

#### **5. K-Nearest Neighbor (KNN)**

KNN method was first introduced by Evelyn Fix and Joseph Hodges in 1951 [23] and later Thomas Cover [24] expended the concept in his research. It is a non-parametric, supervised learning classifier that uses proximity to make classifications or predictions about the grouping of an individual data point. Though it can be used for either regression or classification problems, it is usually used as a classification algorithm. For classification, class label is assigned on the basis of a majority vote whereas for regression problems, the average the  $k$  nearest neighbours is taken to make predictions. In order to decide which data points are closest to a given query point, the distance between two data points will need to be calculated. While, there are several distance measures, Euclidean distance is the most commonly used distance measure.

#### **6. Principal Component Analysis (PCA)**

PCA technique was presented by the mathematician Karl Pearson in 1901 [25]. It is a statistical method that uses an orthogonal transformation to converts a set of correlated variables to a set of uncorrelated variables. The main aim of PCA is to reduce the dimensionality of a dataset while maximum amount of information.

The blue dots show a set of paired data points  $\{(x_i, y_i)\}$ . The PCA applied to this data yields an orthonormal basis shown by the two black arrows. They are ordered by their variance, with principal component 1 (PC1) being the component of most variation. If we then use this to perform a lossy compression we simply discard the principal component 2 (PC2) and perform the inverse transform, yielding the reduced data points shown in green.

#### **7. Decision trees (DT)**

DT classifiers were first introduced by Breiman and his collaborators in 1984 [26]. It is a non-parametric supervised learning algorithm and is used for both classification and regression problems. Decision Tree (DT) Classifier uses a tree representation to answers a given classification question. The decision tree consists of root node, decision nodes, and leaf nodes. Root node has no incoming edges and zero or more outgoing edges whereas decision nodes has one incoming edge and two or more outgoing edge. Class label is assigned to the leaf node that has one incoming edge and no outgoing edge.

#### **8. Random Forest (RF) (Seknedy and S. Fawzi)**

RF introduced by Breiman and Cutler [27], is an ensemble learning method used for classification and regression. Using bagging technique, each decision tree in the ensemble is

built using a sample with replacement from the training data. Each tree acts as a base classifier to determine the class label of instance and class is defined on the basis of majority voting.

## **(ii) Deep Learning Approaches**

### **1. Convolutional Neural Networks (CNNs)**

CNN were presented at the Neural Information Processing Workshop in 1987, automatically analysing time-varying signals by replacing learned multiplication with convolution in time, and established for speech recognition. [28]. CNN extract the feature from the grid-like matrix dataset. It consists of multiple layers like the input layer, Convolutional layer, pooling layer, and fully connected layers. The architecture of CNN is also known as convnets which is a sequence of layers, and every layer transforms one volume to another through a differentiable function.

### **2. Recurrent Neural Networks (RNN)**

The first RNN architecture was introduced by Wilhelm\_Lenz and Ernst Ising [29] [30]. Shun'ichi Amari made the model adaptive in 1972 [31]. RNN uses sequential data to solve common temporal problem in speech recognition. The generalization ability of RNN can be enhanced by attention mechanism and is usually used in speech recognition. It provides an attractive framework for propagating information over a sequence using a continuous valued hidden layer representation.

### **3. Long Short-Term Memory (LSTM)**

Hochreiter & Schmidhuber produced the LSTM, a kind of recurrent neural network that can learn order dependence and output of the previous step is fed as input in the current step [32]. LSTM uses CNN for feature extraction, and compare different dimensions and layers using LSTM to classify emotional features, that efficiently processes signals to increase the performance of voice emotion recognition.

## **III. APPLICATIONS OF SPEECH EMOTION RECOGNITION**

The applications of SER are quite varied and continually growing. Speaker's Speech Emotion recognition could be useful in the field of mental health analysis, psychological disease analysis, stress monitoring, social media analysis, autonomous vehicles, computer game analysis, web based e-learning, online learning to predict student satisfaction, online interviews by analysing their audio or video responses during interviews, marketing, consumer behaviour, customer satisfaction, to address customer grievances, to evaluate the performance of existing employees – especially in the call-centre industry where an improper conversation with a customer can be disastrous for the overall company performance and many more.

Some other areas of application include access control, transaction authentication, law enforcement, speech data management and personalization. Access Control, initially used for physical facilities, is now being applied as biometric factor to usual password or token, computer networks access control through sharing of password to access the subscription sites and automated password reset services. Transaction Authentication are being used for account access control, telephone banking and for more responsive transaction's verification of higher levels. The "e- and m-commerce" verifications i.e. remote electronic and mobile purchases are recent applications. In law Enforcement the applications are applied for prison call monitoring, home-parole monitoring and automatic systems to confirm auditory/spectral inspections of

speech samples for forensic analysis. Since 2008, intercepted speech communication can be accepted as one of the legal pieces of evidence in Indonesian court. Speech Data Management include intelligent answering machines or voice mail browsing to use speaker identification to label incoming voice mail with speaker's name for personal reply.

Emotion, usually reflected in speech, is a reaction that human give in response to an event. Speech emotion recognition technology plays an important role in the field of Human-Computer Interaction. With the technological advancement, AI techniques have been developed in preference to traditional techniques in speech emotion recognition. New-age technologies are being applied by researchers in the field of voice technology to transform their applications, and speech emotion recognition is an important device in this transformation. In this chapter an attempt has been made to explore the machine learning techniques in Speech Emotion Recognition that has been playing significant role in the journey of speech technology.

## REFERENCES

1. PwC, Evolution of voice technology- the next revolution in user interaction. (2022). <https://www.pwc.in/assets/pdfs/consulting/technology/intelligent-automation/evolution-of-voice-technology.pdf>
2. R. Ullah, M. Asif, W.A Shah et al. Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. *Sensors* 23, 6212. 2023. <https://doi.org/10.3390/s23136212>
3. M.B. Akçay, and K. Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, *Speech Communication*, 116, pp. 56-76. 2020.
4. S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.
5. M. El. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
6. S. A. Alim and N. K. A. Rashid "Some commonly used speech feature extraction algorithms" *Natural to artificial intelligence*, Ricardo Lopez-Ruiz (Ed),doi: 10.5772/intechopen.80419
7. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A Review," in *IEEE Access*, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
8. A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *IEEE Trans. neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1421–1432, 2014.
9. L. Deng and D. Yu, "Deep learning: Methods and applications", *Foundation and Trends in Signal Processing* , vol. 7, no. 3, pp. 197-387, 2014.
10. S. Langari, H. Marvi and M. Zahedi, Efficient speech emotion recognition using modified feature extraction, *Informatics in Medicine Unlocked*, vol. 20, 2020. <https://doi.org/10.1016/j.imu.2020.100424>.
11. J.H.Yeh, T.L. Pao, C.Y. Lin et al. Segment-based emotion recognition from continuous Mandarin Chinese speech. *Computers in Human Behaviour*, vol. 27, no.5, pp. 1545-1552, 2011.

12. R. M. Gary, "A history of real-time digital speech on packet networks: Part II of Linear predictive coding and the internet protocol". *Foundation and Trends in Signal Processing*, vol. 3. No. 4, pp. 203–303, 2010. doi: 10.1561/20000000036. ISSN 1932-8346.
13. S.B. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
14. Y. Han, G. Wang and Y. Yang, "Speech emotion recognition based on MFCC", *Journal of Chong Qing University of Posts and Telecommunications* (Natural Science Edition), vol. 20 no.5, 2008.
15. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, 1990.
16. Hönig, Florian, et al. "Revising perceptual linear prediction (PLP)." *Proceedings of INTERSPEECH*. 9<sup>th</sup> European conference on speech communication technology, Lisbon, Portugal, 2005. doi:10.21437/Interspeech.2005-138
17. M. El Ayadi, M.S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, vol.44 no. 3, pp. 572-587, 2011.
18. J. Pearl (1985). Bayesian Networks: A model of self-activated memory for evidential reasoning (UCLA Technical Report CSD-850017). "Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine", pp. 329–334, 1985.
19. J. Pearl, Probabilistic Reasoning in Intelligent Systems. San Francisco CA: Morgan Kaufmann. 1988. ISBN 978-1-55860-479-7.
20. R. E. Neapolitan, Probabilistic reasoning in expert systems: theory and algorithms. 1989, Wiley. ISBN 978-0-471-61840-9
21. C. Cortes and V. Vapnik, Support- vector networks", *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995. Doi:10.1007/BF00994018
22. L.E. Baum, J.A. Egon, "An inequality with applications to statistical estimation for probabilistic functions of a markov process and to a model for ecology". *Bull. Am. Meteorol. Soc.* Vol. 73, pp. 360–363, 1967.
23. E. Fix, J.L. Hodges and L. Joseph L. Discriminatory Analysis. Nonparametric discrimination: consistency properties (Report). USAF School of Aviation Medicine, Randolph Field, Texas (1951).
24. T. M. Cover and P. E. Hart, "Nearest neighbour pattern classification", *IEEE Transactions on Information Theory*, vol. 13, no, 1, pp. 21–27. 1967, CiteSeerX 10.1.1.68.2616. doi:10.1109/TIT.1967.1053964
25. K. Pearson, "On lines and planes of closest fit to systems of points in space". *Philosophical Magazine* Vol. 2, no, 11 pp. 559–572, 2010. doi:10.1080/14786440109462720
26. L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, "Classification and regression trees". Wadsworth, Belmont, 1984.
27. L Breiman, and A. Cutler, "A deterministic algorithm for global optimization". *Mathematical Programming*, vol. 58, no. 1-3, pp. 179-199, 1984. <https://doi.org/10.1007/BF01581266>

28. T. Homma, Atlas, L. and M. Robert, ‘An artificial neural network for spatio-temporal bipolar pattern: application to phoneme classification’. *Advances in neural information processing system*. Vol. 1, pp. 31-40, 1988
29. W. Lenz, “Beitrage zum Verstandnis der Magnetischen Eigenschaften in Festen Korpern”, *Physikalische Zeitschrift* vol. 21 pp. 613-615, 1920.
30. Ising, E. "Beitrag zur Theorie des Ferromagnetismus", *Z. Phys.*, vol. 31 no.1, pp. 253–258, 1925. doi:10.1007/BF02980577
31. J. Schmidhuber, "Annotated history of modern AI and deep learning". *AarXiv*, vol. abs/2212.11279, 2022.
32. S. Hochreiter, and J. Schmidhuber, “Long short-term memory”. *Neural Computation*. vol. 9, no.8, pp. 1735–178, 1997