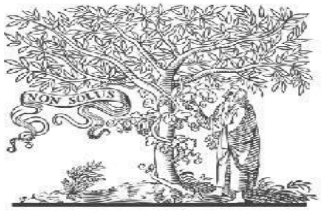
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 21st Feb 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 03

10.48047/IJIEMR/V12/ISSUE 03/21

Title **SPEECH EMOTION DETECTION THROUGH MACHINE LEARNING**

Volume 12, ISSUE 03, Pages: 156-164

Paper Authors
Mrs. Keerthi. G, Madhavi. N, Sowmya. L, Yamini Priyanka. N, Karthik Venkata Kumar. M

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Speech Emotion Detection through Machine Learning

Mrs. Keerthi. G[1], Assistant Professor, Department of Computer Science,
Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru.
keerthi.guttikonda@gmail.com
Madhavi. N[2], Department of Computer Science,
Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru.
Sowmya. L[3], Department of Computer Science,
Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru.
Yamini Priyanka. N[4], Department of Computer Science,
Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru.
Karthik Venkata Kumar. M[5], Department of Computer Science,
Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru.

Abstract
Speech will be most prominent way for humans to communicate with each other. Speech emotion recognition is the procedure of accurately guessing a person's emotion based on his speech. Although it is tricky to annotate audio and difficult to forecast a person's sentiment because emotions are subjective, "Speech Emotion Recognition (SER)" makes this possible. Various researchers have created a variety of systems to extract emotions from the speech stream. Speech qualities in particular are more helpful in identifying between various emotions, and if they are unclear, this is the cause of how challenging it is to identify emotion from a speaker's speech. A variety of datasets for speech emotions, their modeling, and types are accessible, and they aid in determining the style of speech. After feature extraction, the classification of speech emotions is a crucial component, so in this system proposal, we introduced Artificial Neural Networks (ANN model) that are utilized to distinguish sensation like cheerful, disgust, surprise, anger, sad, fear. The proposed system model Artificial Neural Networks (ANN model) achieved precision of training 100% and precision of validation 99%.
Keywords: SER, feature extraction, Artificial Neural Networks Model

Introduction

The majority of modern civilization is powered by machine learning, including social network content filtering, e-commerce website suggestions, and a growing number of consumer items like smartphones and digital cameras. Machine-learning algorithms are helpful to choose suitable search results, recognize objects in photos, convert from speech to text, match newspaper articles, posts, or products with users' interests, and more. These applications are increasingly using a group of methods known as deep learning. For many years, building machine learning or pattern recognition frameworks have required experience and a lot of domain knowledge from developers to develop feature extraction methods that transform raw data into suitable input files or feature vectors. The training subsystem, which is primarily a classifier, can detect or classify patterns in the input data. A set of techniques known as “representational learning” allows computers to receive unstructured data and immediately discover patterns needed for classification or identification.

TESS: Toronto Emotional Speech: This dataset consists of similar kinds of sentences in seven different emotions (happy, disgust, surprise, angry, sad, fear, and neutral) entirely equal to 2800 audio samplings.

Literature Survey

Depth effort has been performed in recent years to determine feelings through speech records. In this application, SVM algorithms are directed for precise feelings, deciphering the data from each character in

a query separately to synthesize emotion detection records for addressing binary category problems, and then Multi-class prediction is applied by combining forecasts from rankers. Ranking SVM has a number of advantages, Including obtaining speaker-specific training and testing data.Second, To determine the dominant emotion, it takes into account the knowledge that each speaker may display a variety of emotions.
New et-al. [8] suggested a new technique to recognizing emotions in speech indicators. To characterize the speech signals and classifier, the system used LFPC which means a log-frequency power coefficient and a seperate HMM. This strategy will divided emotions into six separate groups before training and testing the new system with private information. To assess the effectiveness of the suggested method, log-frequency power coefficient be the comparision for MFCC and LPCC. This outcome shows that standard accuracy 78%, finest categorization accuracy 96%, accordingly. Additionally, the outcome show that higher-level option Will be LFPC as an emotion categorization property to standard characteristics [8].

Wu et al. [6] suggested modern "MFS" which means modulation spectral features for feeling identification in human speech. Adequate features were retrieved out of an auditory-stimulated long-term spectro-temporal with the usage of an attenuation filter out bank and an auditory clear out bank for voice deconstruction. This method has given auditory frequentness and secular modulation frequency elements to transmit critical data that typical Temporal spectral characteristics. SVM with radial basis function (RBF) is used in the classification process. MSFs are evaluated in "VAM" which means Vera am Mittag and Green city(Berlin). According to the experimental results, Modulation spectral features outperform Mel-frequency Ceptral coefficients and "PLPC" which means perceptual linear prediction coefficients. When MSF are used to augment prosodic characteristics, recognition performance improves significantly. Furthermore, for categorization, a 91.6% total recognition rate was reached.

Rong's et-al. [10] developed "ERFTrees" known as extended trees at random forest approach by plethora with traits to feeling reputation with out regarding any linguistic records, but, the hassle of linguistic records stays unsolved. This method is used on tiny amounts of data with a big range of capabilities. An test using a chinese language emotional speech dataset changed into used to check the advised technique, and the findings show that this approach improved on emotion reputation fee. moreover, ERFTrees surpass preferred measurement discount procedures like "MDS" which is known as multi-dimensional and "PCA", as well as the currently delivered ISO.Great percent of 16 traits towards a woman records received a most precise percent of 82.54%, while the worst accuracy with 84 functions completed best 17%.

Albornoz et al. [16] take a look at a brand new spectral belongings that can be used to assess emotions and describe corporations. emotions are categorised on this work the use of auditory traits and a completely unique hierarchical classifier. one of a kind classifiers, together with GMM, HMM, and

MLP, had been investigated with one-of-a-kind configurations and enter data to construct a brand-new hierarchical system for categorising emotions. The suggested approach is accurate in that it first chooses the capabilities that perform well, and then it uses the great elegance-wise classification performance of all capabilities, including the classifier. The results of the experiments conducted on the Berlin dataset demonstrate that the hierarchical approach outperforms the top-performing general classifier with multiple go-validation. For instance, the hierarchical model achieved 71.57% while the conventional HMM technique achieved 68.57% performance.

Yeh et al. [9] advised an emotion identification technique based totally on

segmentation in Mandarin speech. the subsequent manner is protected on this technique.first, give the stacked discrete k-nn classifier the k parameter. Many k values are tested, but when k is set to 10, the performance for k-nn is found to be at its best. The most crucial feature set is chosen using sequential ahead selection (SFS) and sequential backward choosing (SBS). When SBS and SFS are used, feature accuracy will rise to 82% and 84%, respectively. The strategy with the best accuracy, 86%, is segment-based. By luring in 18 men and 16 women, a private corpus was used to support the experimental findings. In order to advance the extensive emotional inquiry, more expressive speech must be gathered. [9].

El Ayadi-al-et. [3] a mixture of GMM and vector autoregressive called Gaussian aggregate vector autoregressive (GMVAR) solution for the category issue of voice emotion recognition. The core concept of GMVAR is its capacity to distribute information across several media and foster trust among speech feature units. Using the Berlin emotional dataset, GMVAR is assessed. According to the experimental findings, feed-forward neural networks have a category accuracy of 55%, 71% for k-NN, and 76% for HMM. For neutral emotions, this technique has a bigger distinction between high and low arousal in comparison to HMM [3].

Arias's at-et. [11] introduced a unique shape-primarily based technique for detecting emotional salience in the essential frequency by employing a neutral model. the radical method, that's supported with the aid of practical records analysis (FDA), aims to acquire the natural variability of F0 contours. PCA is generated for a specific F0 contour for use as a characteristic for speech emotion popularity. The empirical effects display that the proposed method acquired 75.7% accuracy in binary category. It equates to a 6.3% improvement over the trained benchmark system with general F0 statics. The SEMAINE dataset is used to assess the strategy. The findings imply that the use of the shape-based totally approach in real-international applications to decide speech emotion may be useful facial expression popularity has garnered quite a few attention in social technology and computer-human interplay. Profits from deep learning have produced advancements in this field that go beyond precision at the level of a person. The object[2] has looked at and favours deep learning algorithms for recognising emotions, while also using the eXnet library to boost precision. On the other hand, memory and computing continue to be challenging tasks. Overfitting can happen to large models. Reduce generalisation errors as one approach to this problem. We build a novel CNN version that utilises concurrent function extraction using a single Convolutional Neural network (CNN) called eXnet. The most recent "eXnet" version of Expression Net uses far less parameters while improving upon the older version's accuracy.

Deaf and dumb people all over the world use Sign Language to communicate. However, communication between a verbally handicapped person and a normal person has always been challenging. Sign Language Recognition is a significant advancement in assisting deaf-mute individuals to communicate with others. Today, academics all over the world are concerned with commercializing an affordable and accurate recognition system. As a result, sign language recognition systems that utilize image processing and machine learning are chosen over gadget systems because they are more accurate and simpler to implement. The Aim of the study[1] for creation of a user-friendly and accurate recognition of a sign language system that is trained using a neural network and can generate text and voice from the input motion.

Emotion identification from speech has recently gained popularity among researchers. The study[5] covers numerous methods for recognizing emotions in audio signals that employ K-NN, random forest and multiple level perceptron of cnn which are machine learning algorithm, Random

forest, multi-layer perceptron, and convolutional neural network. This emotional speech database was used to generate short-term Fourier transform spectrograms and MFCC. Spectrograms were fed into CNN as input. MFCC properties were fed into k-NN, MLP, and random forest

models. Each classifier correctly classified seven emotions (happy, unhappy, irritated, impartial, disgust, boredom, and worry), but the MLP classification stood out with an average accuracy of 90.36%. A comparison of the overall performance of those categorization algorithms is also provided.

This study suggests using a "GPLDA" back-end, or Gaussian Conditional Linear Discriminant Analysis, to classify emotions at the syllable level using "I"vectors that encode the distribution of frame-stage MFCC capabilities. The GPLDA back-end beats an SVM-based back-end despite being less sensitive to the i-vector measurement, according to the results of the IEMOCAP corpus experiment, making the suggested framework more resistant to variable tuning throughout device development[4] presented a reconstruction-errors-primarily based (RE-based) getting to know framework with reminiscence-greater Recurrent Neural Networks to improve the overall performance of continuous emotion reputation from speech (RNN). The system employs successive RNN models, the primary serving as an autoencoder for reconstructing the authentic functions and the second for emotion prediction. The authentic features' RE is applied as a complimentary description, that is then mixed with the original functions and positioned into the second one model. The framework assumes that the system has the ability to study its 'downside,' which is expressed by way of the RE. In terms of overall performance, experimental effects revealed that the proposed architecture significantly outperforms baseline systems without RE records on the RECOLA collection.

Existing System

Gamage et al. provided Gaussian analyses to separate conversation feeling ranges based on "I" vectors, proving the efficacy of MFCC dispersion. The GPLDA basis surpasses the SVM basis and is less sensitive to the I -vector, according to an analysis based entirely on the IEMOCAP corpus. As a result, the bottom rate makes it more effective to change laws at a later stage of framework development. Defending Han et alproposals's for an enterprise that recursively develops memory and a teaching error-based approach to handle learning. Due to this, two continuous RNN (Recurrent Neural Networks) techniques are used, with the first version serving as an automated code to retrieve the preliminary data and the second version serving as a passionate forecast. When employed as a secondary aid, the primary aid's RE (Reconstruction-blunders-based) is closely matched to the first aid and positioned in the second elegance. SVM is used in the contemporary machine to apprehend the speaker's feelings. assist Vector machine (SVM) is a kind of supervised device gaining knowledge of that has been used for category and regression. But this existing system has few disadvantages like Support vector Machines don't function well in the current system model because classes are overlapped. Within the current system paradigm, it is difficult to choose the appropriate kernel for the SVM. On a huge data collection, the current system model requires more time to be trained. Support vector machines, or SVMs, are not deterministic models, hence it is difficult to detail the classification using statistics. SVM is more sophisticated than a Decision tree, making it more challenging to comprehend and analyze. When the desired groups are coinciding and the data set includes more uncertainty or noise, SVM doesn't function very well.

Proposed System

The goal of computing is to enable efficient and natural human-computer interaction. One crucial objective is to make it possible for computers to comprehend the emotional states that people express so that tailored

responses may be given. The majority of research in the literature concentrates exclusively on recognizing feelings from short, isolated words, which prevents practical applications. We use artificial neural networks to implement voice emotion recognition in the suggested system (ANN model). The proposed system, which comprises seven different emotion categories, is based on experiments using pre-recorded datasets that were carried out by Kaggle. The proposed system achieves a target train accuracy of 100% and the target test accuracy of 99%.

The system requests training data, which includes weight training and expression labeling for that network. The input is an audio file. The audio is then subjected to intensity normalization. To prevent the impact of the presentation sequence of the samples from affecting the training performance, the ANN is trained using normalized audio. The weight collections that are produced as a result of this training procedure produce the best outcomes when used with the learning data. Using the final network weights learned, this dataset, which was used for testing, retrieves the system with pitch and energy and offers the detected emotion. This proposed system has many advantages like Artificial neural networks can give data for parallelization, allowing them to tackle multiple tasks at once. Even in the absence of a data pair, the network is still capable of producing results since ANN is used to store data. Resistance has existed in them which implies that the functionality of ANN is affected when one or more neural networks are lost. Artificial neural networks are progressively dissipating, so they won't instantly stop functioning. Because of this, these system is gradually dissipating. Ability to teach Artificial neural networks that these networks learn from previous events to conclude.

Artificial neural networks

Similar to the extensive network of neurons located inside the mind, an artificial neural network is made up of interrelated nodes in groups. Each circular node indicates an artificial neuron, and each arrow shows how one artificial neuron's output connects to another's input. Computer systems called artificial neural networks (ANNs) are loosely modeled after the biological neural networks that make up animal brains. Such kinds of systems "learn" to execute tasks by taking into account examples, typically without having any endeavor rules written into them.

Feature extraction and model training are necessary for emotion detection from speech data. The feature vector is made up of audio signal elements that identify speaker-specific properties such as energy, amplitude, and tone, and it is used to teach the classifier model to identify a specific emotion precisely. TESS open source dataset which is in the English language included both female and male speakers' different acted speech corpus and was manually separated into testing and training phases. Mel-frequency cepstral coefficients reflect speaker vocal tract information, and we recovered the MFCC coefficients from the audio samplings in the TESS dataset. On collected human speech, we also used feature extraction. The energy and MFCC coefficients of several emotion aural recordings, such as calm, anger, anxiety, and melancholy, were determined.
“A-NN” was the algorithm that is composed by a connection spot occurred in artificial neural network known as "artificial neurons," that roughly model neurons placed in human mind. Individual link, like neurons in the actual brain, can send information, or a "signal," from one biological neuron to the next. When the neuron in the ANN received a gesture, it will estimate the signal then signal remaining neurons that are linked to it. The wave at the link in between neurons is a real number in most artificial neural network implementations, as well as the outcome of every neuron, is generated through certain arbitrary expressions by aggregate sources. Edges are the connectors between biological neurons. The density of artificial neurons and edges is often adjusted as learning progresses.

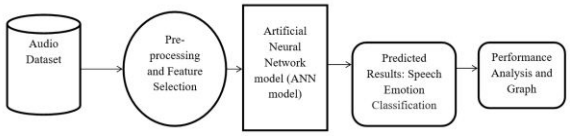
Artificial neurons are typically organized into layers. Each layer has its significance. Each layer may apply various transformations to its inputs. Signals go to the final layer (outcome layer) of first level (input sheet), sometimes many times. Then weight changes the quality of the signal at a junction. Artificial neurons may possess a boundary that is crossed by the aggregate signal before the signal is transmitted.

The ANN approach's primary intent was always to solve issues in the same manner that a human psyche would. Nevertheless, as time passed, the focus shifted to certain tasks, resulting in departures from biology. Artificial neural networks were employed to perform a wide range of tasks, involving feature extraction, voice recognition, language processing, social media network screening, board and games console play, and medical issue.

Speech emotion recognition system and how it works Translating speech to text was the first step in speech recognition research. The initial amount of data was therefore gathered. In advanced technological implementations, the environment and commiserating with the presenter become critical for recognizing speech emotions. Text sentiment classification differs from spoken emotion recognition in this regard. The emotion is represented explicitly in the context of sentiment analysis, making it easier to understand the original intent. In SER, meanwhile, this entire content is hidden beneath the initial layer of data.

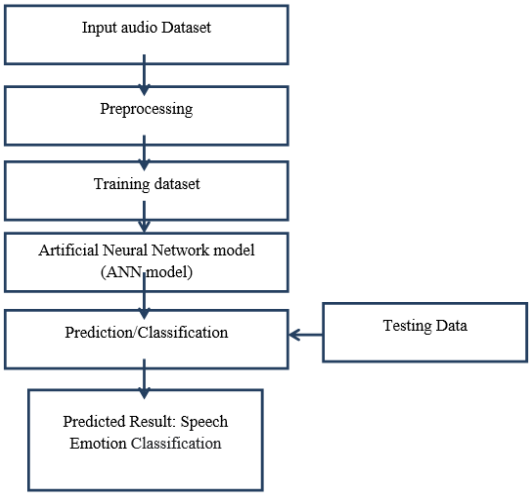
Multiple audio analysis methods are used by scientists to collect this underlying piece of information, which can magnify and recover tones and auditory elements from voice. Translating acoustic signals to numerical or vector format is more difficult than converting graphics. When we forsake the "sound" format, the translation procedure will decide how much critical information is maintained. If a specific data conversion is unable to preserve tenderness and serenity, it will be difficult for the algorithms to identify the sample and learn the feeling. Mel Spectrograms, which depict audio signals depending on their spectral analysis and may be projected as an auditory wave and supplied to train an ANN as a visual classifier, is one approach for converting audio data to numeric. Mel-frequency cepstral coefficients can be used to capture this (MFCCs). Each of these file formats offers advantages and downsides depending on the application.

System Architecture



Data Flow Diagram (DFD)

DFD demonstrates the information's flow through the system and the various changes that affect it. The following data flow diagram depicts the information flow and the modifications made to data as it travels from the input audio dataset to the output result.

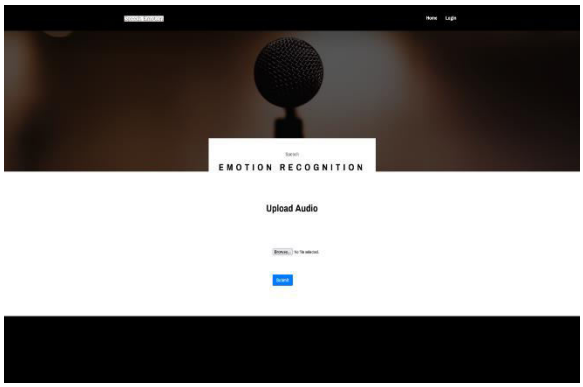


International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

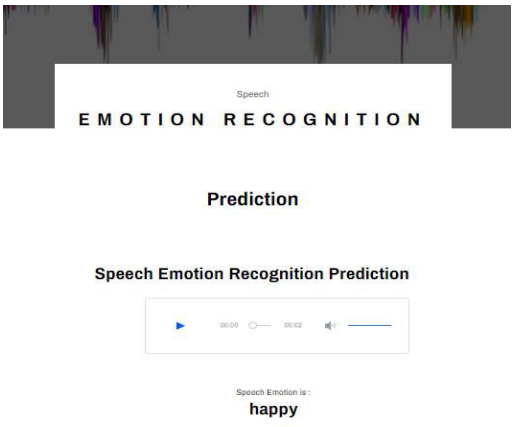
Experimental Results

After successful installation of required packages and libraries, the python file gets executed which is followed by the output screen as shown

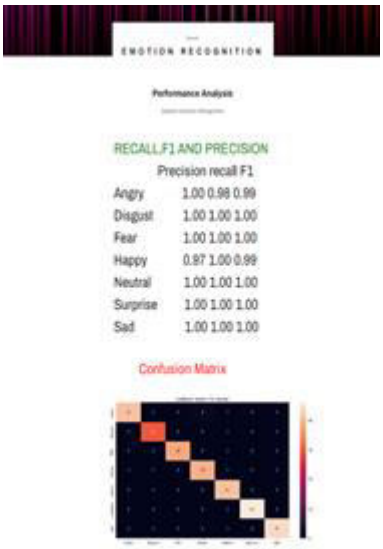
Valid credentials are given which is directed to the page which contains accepting the audio file. An audio file from the TESS dataset is uploaded as follows



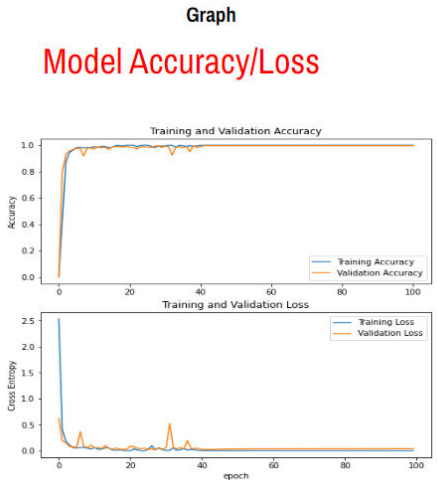
The emotion of the audio is identified and it is shown on the screen itself as follows.



The overall performance of this model is given as follows



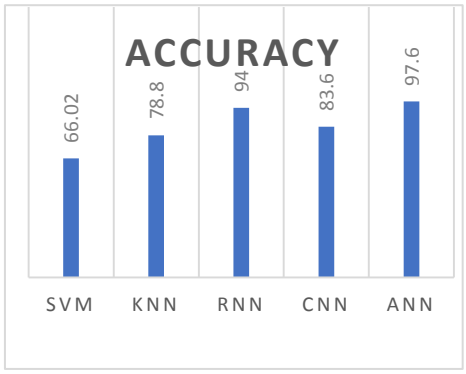
The model accuracy/loss graph is depicted as follows



ANN algorithm is compared with different other algorithms it is clear that its accuracy is better than others as follows

Conclusion and Future Scope

Systems for recognizing speech emotions based on several classifiers are shown. The signal processing unit in a voice emotion recognition system is one of the key concerns.



The appropriate features are extracted from the available speech signal by a classifier that distinguishes emotions from the speech signal.Most classifiers' average accuracy for speaker-independent systems is lower than that for speaker-dependent systems.

Automatic emotion detection from the human speech is becoming more prevalent today because it improves interactions between humans and machines. Combinations of the techniques mentioned above can be developed to enhance the emotion recognition process. Additionally, By removing more helpful speech components, the speech emotion detection system's accuracy can be increased.

The suggested architecture might further be expanded to provide multiple languages of Emotion. Feelings can also be used to describe small aspects and design.

References

[1] Thaikur, Shirish Shrestha, Sarmila Upreti, Amrita, and Subarna Shakya, Pujan Budhathoki investigated Real-Time Sign Language Recognition and Voice Generation.

[2] Koty ursami and Koti ilingame published a review on developing an efficient method for detecting customer emotion analysis through deep learning analysis.

[3] M. H. El Ayady, M. S. Kamil, and F. Karrey, "Speech Emotion Recognition Using Gausyan Mixture Vector Autoregressive Models," in 2007 IEEE International Conference on Acoustics, Voice, and Signal Processing - ICASSP '07, vol. 4, pp. IV-957-IV-960, 2007.

[4] Jingy Hain, Zixeng Zhong, Febien Ryngeval, and Bgorn Schuler investigated reconstruction-error-based learning in speech for continuous emotion recognition.

[5] Speech Emotion Recognition Using CNN, k-NN, MLP, and Random Forest was investigated by Kauri, Jasmeeti, and Anyl Kumar.Kalany Wataraki Gamages, Vidhyasaharran Sathu, Phu Ngoc Le, and Eliathamby Ambikayrajah conducted research on an i-vector GPLDA system for speech-based emotion recognition.

[6] "Automatic speech emotion recognition through modulation spectral features," YS. Wu, Ti. H. Falk, and W.V. Chan, Speech Conversation., vol. 53, no. 5, pp. 768-785, May 2011.

[7] "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," H. Ceo, R. Varma, and A. Nankova, Compute. Speech Lang., vol. 28, no. 1, pp. 186-202, Jan. 2015.

[8] "Speech emotion recognition through hidden Markov models,"DeSilva.L.C,Foo.S.W,New.T.L, Speech Commun., vol. 41, no. 4, pp. 603-623, Nov. 2003.

[9] "Segment-based emotion recognition from continuous Mandarin Chinese voice," G.H. Yeh, Y.L. Pao, K.Y. Lin, F.W. Tsai, and Y. Chen, Comput. Human Behav., vol. 27, no. 5, pp. 1545-1552, Sep. 2011.

[10] "Acoustic feature selection for automatic emotion recognition from speech," G. Rong, J. Lye, and W.P. Chen, Info. Procedure. Manager., vol. 45, no. 3, pp. 315-328, May 2009.

[11] "Spoken emotion recognition through hierarchical classifiers," Y.-M. Albornoz, Di.-H. Milone, and Henry Rufiner, Comput.

Speech Lang., vol. 25, no. 3, pp. 556-570, Jul. 2011.

[12] J. Pi. Arias's, Ci. Buss, and N.-B. Yomas, "Shape-based modelling of the fundamental frequency contour for emotion detection in voice," Comput. Speech Lang., vol. 28, no. 1, Jan. 2014, pp. 278-294.