**Literature Review (Secondary Research) Template**

|  |  |
| --- | --- |
| **Student Name** | **Vasamsetty Lohitha** |
| **Project Topic Title** | **SpeechSentio-an AI powered speech therapy with emotion analysis** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of Variables that You Need to Search for in Each Article (Each Current Solution)** | | | |
| **Dependent variable** | **Independent variable** | **Moderating variable** | **Mediating ( Intervening) variable** |
| * The presumed **effect** in an experimental study. * The values of those variable depend upon another variable that are the independent variables. * Strictly speaking, “dependent variable” should not be used when writing about non-experimental designs. | * The presumed **cause** in an experimental study. * The variables that may impact on the dependent variable * The values of those variable are under experimenter control. * Strictly speaking, “independent variable” should not be used when writing about non-experimental designs. | * has a strong  *contingent*effect on the independent variable-dependent variable **relationship** and thus produces an interaction effect. | * It comes between the independent and dependent variables and shows the **link or mechanism** between them. |
| * Examples: **1.** **performance**. **2.** **Test Score**. **3.** **stock market. 4. performance** of the  students | * Examples: **1.** **run time** that will impact and cause high/low performance. **2.** **Time Spent Studying** that will cause the high/low score. **3.** **New product**  that will impact on the  stock market price. **4.** **quality of  library facilities** | * Example: **4.** There is a strong relationship  between the quality of  library facilities  (X) and the performance of the  students  (Y). Only  those students who have the **interest and  inclination** to use the  library  will show improved performance in their studies, which moderates the strength of the association between X and Y variables. | * Example: Parents transmit their social status to their children directly, but they also do so indirectly, through education: viz. Parent’s status ➛ child’s education ➛ child’s status * Example: The statistical association between income and longevity needs to be explained because just having money does not make one live longer. Other variables intervene between money and long life. People with high incomes tend to have better medical care than those with low incomes. Medical care is an intervening variable. It mediates the relation between income and longevity. |

|  |
| --- |
| **Relationship among Variables - Correlations (Univariate, Bivariate, Multivariate)** |
| * Once the variables relevant to the topic of research have been identified, then the researcher is interested in the relationship among them. * A statement containing the variable is called a **proposition**. It may contain one or more than one variable. * The proposition having one variable in it may be called as **univariate**  proposition,  those with two  variables as **bivariate**  proposition, and then of course  **multivariate** containing  three or more variables. * Prior to the formulation of a proposition the researcher has to develop strong  logical arguments  which could help  in establishing the  relationship. * For example, age at marriage and education are the two variables that could lead to a proposition: the higher the education, the higher the age at marriage .  What could be the logic to reach this conclusion? All relationships have to be explained with strong logical arguments. If the relationship refers to an observable  reality, then the proposition can be put to test, and any testable proposition is hypothesis. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Research Model That The Author Followed to Propose His Solution** | | | |
| **1. Where we are now** | **2. Where are we going** | **3. How do we get there** | **4. How do we know when we are finished** |
| * What the author has done in the area; The constructs that the literature examine * **What the problem is available** in this paper that has solved by the author * The purpose of that is to avoid pursing research which has already been undertaken | * What the author **objective** of the research is to gain a clearer understanding the relationships between variables * What is the goal of the paper * The purpose is to know what is the plan to do before he did the research | * How the author conducted the research; **How the problem has solved** * How he analysed the data generated by the research; A quantitative research design | * What is the value of this solution * A series of **recommendations** which flow from the data analysis have been made |

**NOTE: Please you need to use YOUR OWN WORDS in writing this template.**

**Your Literature Review Should be in Scope and MUST Address all Your Project's Questions**

|  |
| --- |
| **1** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| https://www.mdpi.com/1424-8220/20/1/183 | | Mustaqeem  Soonil Kwon | | | | Artificial Intelligence, emotion recognition,convolutional neural networks (CNN), noise removal, spectrogram, signals enhancement | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| An artificial intelligence-assisted deep stride convolutional neural network (DSCNN) | | Increasing the accuracy of speech emotion recognition and reducing the computational complexity of the model. | | | | The system includes three components:  1.Audio signal preprocessing for spectrogram generation.  2. Deep stride convolutional neural network (DSCNN) architecture for extracting deep features from the spectrograms and classify the emotions.  3. post-processing, involves in decoding the predicted emotion labels and generating the final output. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Data is pre-processed to remove noisy and silent portions. IEMOCAP and RAVDES datasets are used. | The model has a training time of only 14 minutes, demonstrating its computational efficiency compared to other state-of-the-art CNN models | Computational cost is more to train and run. | | **2** | Spectrograms are created from the speech data and are fed into a CNN model to extract features. | The model has a smaller model size compared to other state-of-the-art CNN models, making it computationally simpler | The proposed model may require significant computational resources for training on larger datasets. | | **3** | Features are used to classify emotion of the speaker. | The proposed CNN model improves the overall prediction accuracy, indicating its robustness | The model may not perform well on noisy data. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Speech emotion recognition accuracy | Deep Stride Convolutional Neural Network | Specific features learned in the DSCNN | Characteristics of datasets | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | |  | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Audio signals | Speaker emotion | | | | The system features audio signal preprocessing, a DSCNN architecture for feature extraction, and emotion classification, followed by post-processing for generating the final output. | | | | DSCNN model and the adaptive threshold-based pre-processing of the speech signal improved the accuracy of speech emotion. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| Accuracy, identify user emotional state for appropriate response and selection of the classifiers with outlier detection enhances the speech emotion recognition process. | | | | | Oversimplification of human emotions, misinterpretation of emotions. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The proposed method first enhances the audio signal using a novel method based on Convolutional neural networks, and then classifies the emotions using SVM. The method is novel and effective and the experimental results are convincing. | | | | Convolutional neural network (CNN), Deep Neural Network (DNN) | | | Abstract   1. Introduction 2. Related Work 3. Proposed Methodology 4. Experiments and Results 5. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

**---End of Paper 1-**

|  |
| --- |
| **2** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| https://ieeexplore.ieee.org/document/8070805 | | Mohan Ghai  Shamit Lal  Shivam Duggal  Shrey Manik | | | | Berlin database, Emotion recognition, Gradient boosting, MFCC, SVM, Random Forest | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Random Forest | | The goal of the paper is to recognize emotions in speech signals and classify them into seven emotion output classes using machine learning techniques. | | | | Author used Mel Frequency Cepstral Coefficients (MFCC) and Berlin database of emotional speech. The components of the paper also included feature vector, classifiers which are used to recognize emotions in speech signals. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Feature Extraction | High accuracy in recognizing emotions in speech signals. | Only limited to 7 human emotions. | | **2** | Three Classification Algorithms are used for classifying an audio signal into one of the 7 classes. | Effective use of Mel Frequency Cepstral coefficients and energy of speech signals as feature inputs. | The classification algorithms wrongly predicted some of the samples belonging to happiness class as belonging to anger class. | | **3** | Results for each algorithm is summarized. | Potential for diverse applications in the field of interaction between humans and computers. | Increased computational complexity and risk of overfitting. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Emotion classification accuracy | MFCC, Energy of speech signals | Characteristics of the Berlin database of emotional speech | The process through which Mel Frequency Cepstral coefficients (MFCC) and energy of speech signals influence the relationship between the classification algorithms | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | |  | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Emotional speech | Detected Emotion | | | | The system leverages perceptual features including Mel Frequency Cepstral coefficients (MFCC), energy of the speech signals, and the classification algorithms Support Vector Machine (SVM), Random Decision Forest, and Gradient Boosting. | | | | Considering the different classification strategies, the maximum accuracy is obtained for the database by using Random Decision Forest classifier. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| Random Forest and Ensemble methods like gradient boosting are good advancement to improve performance and accuracy. | | | | | Speech emotion recognition models are always not accurate. They can be fooled by accents, background noise, and other factors. This could lead to misunderstandings and misinterpretations of people's emotions. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| Logically this is a good step that detects emotions in a speech and tests accuracy. | | | | Support Vector Machine (SVM), Random Forest, Gradient boosting. | | | Abstract   1. Introduction 2. Berlin Database of Emotional Speech 3. Related Work 4. Speech Emotion Recognition Framework 5. Experiment Results 6. Conclusion 7. Future Work |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

**--End of Paper 2—**

|  |
| --- |
| **3** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| https://www.mdpi.com/1424-8220/22/6/2378 | | Apeksha Aggarwal  Akshat Srivastava  Ajay Agarwal  Nidhi Chahal  Dilbag Singh  Abeer Ali Alnuaim  Aseel Alhadlaq  Heung-No Lee | | | | Speech emotion recognition, machine learning, Principal Component Analysis (PCA), deep neural network (DNN), feature extraction. | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Two-way feature extraction | | The goal of the system is to improve speech emotion recognition by exploring two different methods of feature extraction. | | | | The system comprises two-way feature extraction methods, Principal Component Analysis (PCA), Deep Neural Network (DNN) and pre-trained VGG-16 model, multimodal speech data. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Two datasets are used Toronto Emotional Speech Set (TESS), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). | The extracted features are processed using principal component analysis (PCA) and a deep neural network (DNN) with dense and dropout layers. This combination of feature extraction and modelling techniques helps improve the accuracy of speech emotion recognition. | Increase in complexity of the system | | **2** | Two approaches are introduced for extracting features: i) working directly on the audio dataset to obtain numerical features, ii) utilized spectrograms as image features. | By utilizing super convergence, the method extracts two sets of potential features from the speech data. This approach allows for a more comprehensive representation of the emotional content in the speech signals. | Dependency on pre-trained models | | **3** | The VGG-16 model outputs a feature vector. The feature vector is classified into one of the seven emotions. | It also involves multimodal data integration. | Sensitive to input quality and limited interpretability. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Speech emotion recognition accuracy | Two-way feature extraction methods | Dataset used (RAVDESS vs numeric features on a DNN) | Process through which two-way feature extraction methods super convergence. | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | |  | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Raw audio signal | Finding the emotion of the speech | | | | Feature extraction plays a crucial role in the process of recognizing speech emotions. Utilizing PCA for feature extraction with DNN and utilizing pre-trained VGG-16 model for speech emotion recognition. | | | | To the extent this work is designed for the two -way feature extraction method for Speech Emotion Recognition. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The proposed SER method improves feature extraction for emotional communication and excels in accuracy, adaptability, and practicality. | | | | | Negative impact of this solution include dependency on dataset characteristics, regional bias, and varied accuracy across emotions. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| Since this is designed involving Deep learning and feature extraction from both audio and Mel spectrograms which made a valuable addition to the field, the method’s evaluation on a single benchmark dataset raises concerns about its generalizability. | | | | Accuracy, f1 score. | | | Abstract   1. Introduction 2. Materials and Methods 3. Results 4. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

**--End of Paper 3-**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Version 4.0 Week 4** | | | | | | | |
| **4** |
| **Reference in APA format** | |  | | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | | |
| https://ieeexplore.ieee.org/document/9528931 | | Tedd Kourkounakis  Amirhossein Hajavi  Ali Etemad | | | | Stuttering, deep learning, disfluency, Mel-frequency spectral coefficients (MFCC), Bidirectional Long short term memory(BLSTM), Squeeze-and-Excitation residual (SE-ResNet). | | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | | |
| FluentNet | | To detect and classify different forms of stuttering and speech disfluencies. | | | | The system components encompass Squeeze-and-Excitation Residual Convolutional Neural Network, Bidirectional Long Short-Term Memory (LSTM) Layers and Attention mechanism | | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Collecting speech samples from UCLASS and LibriStutter dataset. | Provides End-to-end solution | Limited dataset size, computationally expensive. | | **2** | Generating spectrogram from audio clips. | Spectral Frame-level representations | Challenges in classifying interjections | | **3** | Utilizing a Squeeze-and-Excitation Residual Network (SE-ResNet) and passing spectrogram feature vectors through bidirectional LSTM layers. | Incorporates temporal relationships | Developers of FluentNet do not have a clear privacy policy in place, which raises further concerns about how user data is being used. | | **4** | Training model on annotated dataset and evaluate performance. Analyzing experimental results for accuracy. | Achieves state-of-the-art performance. |  | | **5** |  |  |  | | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Stutter detection and recognition performance | FluentNet architecture, stutter types | Characteristics of datasets | Learning of strong spectral frame-level representations | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | |  | | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** | |
| |  |  | | --- | --- | | **Input** | **Output** | | Short-Time Fourier Transform (STFT) spectrograms of audio clips | Types of stutters (sound repetition, word repetition, prolongation etc) | | | | Features of the system include Squeeze-and-Excitation Residual Network, Bidirectional Long Short-Term Memory (BLSTM) Layers and global attention mechanism for stuttering detection. | | | | This model introduces a novel method of stutter detection and it addresses potential errors and computational complexity associated with ASR. | |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | | |
| The authors found that using this model could potentially be used to develop smart and interactive tools for detection and therapy, as well as to improve presentation skills. | | | | | The FluentNet model could misdiagnose people with stuttering, which could lead to them receiving inappropriate treatment. | | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** | |
| This work innovatively addresses stutter detection by directly analysing audio signals, eliminating reliance on automatic speech recognition. The integration of Squeeze-and-Excitation Residual Network (SE-ResNet) and bidirectional LSTM layers contributes to effective feature learning, providing potential advancements in capturing variety speech patterns. | | | | FluentNet, a deep neural network, and several baseline and state-of-the-art techniques. | | | Abstract   1. Introduction 2. Related Work 3. Proposed Method 4. Experiments 5. Conclusion | |
| **Diagram/Flowchart** | | | | | | | | |
|  | | | | | | | | |

**--End of Paper 4--**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Version 5.0 \_ Week 5** | | | | | | | |
| **5** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| https://arxiv.org/abs/2002.07590 | | Manas Jain  Shruthi Narayan  Pratibha Balaji  Bharath K P  Abhijit Bhowmick  Karthik R  Rajesh Kumar Muthu | | | | Emotion, Support Vector Machine (SVM), Mel Frequency Ceptral Coefficients (MFCC), feature extraction, LPCC. | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Support Vector Machine | | The goal of this work is to address the identification and classification of speech into various emotions. | | | | The systems components include the input speech signal, feature extraction using MFCC and LPC, classification based on SVM, and the output. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | The author used two datasets, LDC and UGA as input. | SVM-based SER system has demonstrated high accuracy in recognizing emotions from speech utterances. | The system’s training had very less amount of labelled speech data. | | **2** | Extracted acoustic and prosodic features from speech recordings and trained an SVM classifier using the selected features and emotion labels. | This system is robust to noise and variations in recording conditions, and can be adapted to different languages and accents. | The model is Computationally expensive. | | **3** | Evaluated the trained SVM classifier on a separate test dataset to assess its performance. |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| <Find all main factors and variables that are related to each solutions. Then find the relationship between factors. (Independent variable) causes a change in (Dependent Variable) and it isn't possible that (Dependent Variable) could cause a change in (Independent Variable).   |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Emotion classification | Acoustic features and prosodic features | Speaker characteristics | Contextual information | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | |  | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Raw audio recording of speech | Emotion Classification | | | | The SVM-based SER system demonstrates high accuracy and robustness in recognizing emotions from speech utterances. Its ability to handle non-linear relationships and adapt to different languages further enhances its versatility. | | | | The SVM-based SER system offers a novel approach to speech emotion recognition, combining support vector machines with acoustic and prosodic feature extraction to achieve high accuracy and robustness in real-world applications. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| This model has improved customer satisfaction by accurately recognizing customer emotions, tracks changes in emotional patterns over time. | | | | | Misinterpretation of emotions could lead to inappropriate responses, misunderstandings, and further emotional distress. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The SVM-based SER system demonstrates promising advancements in speech emotion recognition, exhibiting high accuracy, robustness to noise, and the ability to handle non-linear relationships between features and emotions. However, it is essential to critically evaluate its potential limitations and broader implications. | | | | Confusion matrix, f1-score | | | Abstract   1. Introduction 2. Methodology 3. SVM Algorithm 4. Datasets 5. Simulation outputs and results 6. Conclusion and Future Works |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

**--End of Paper 5—**

**Work Evaluation Table**

**<Use the same factors you have used in "Work Evaluation Table" to build your own “Proposed and Previous comparison table ">**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Work Goal** | **System's Components** | **System's Mechanism** | **Features /Characteristics** | **Cost** | **Speed** | **Security** | **Performance** | **Advantages** | **Limitations /Disadvantages** | **Platform** | **Results** |
| **Mustaqeem and Soonil Kwon(2020)** | To enhance the accuracy of a speech emotion recognition system. | The system consists of three stages: Audio signal processing, Deep stride convolutional neural network (DSCNN) and generating the final output. | The proposed system uses audio signal preprocessing to generate spectrograms, which are then fed into a DSCNN architecture to extract deep features and classify emotions. | Usage of a novel adaptive thresholding technique to remove noise and unimportant portions from speech. |  | The proposed DSCNN model has a training time of 14 minutes. |  | The proposed DSCNN model demonstrates strong performance, achieving an overall accuracy of 81.75% on the testing dataset. | 1. Efficient training time  2.High Accuracy  3.Computational simplicity  4.Effective preprocessing | 1.Limited interpretability  2.Sensitivity to Hyperparameter Tuning  3.Generalization Challenges |  | The proposed model achieves an overall accuracy of 81.75% on the testing dataset, outperforming other state-of-the-art CNN model |
| **Mohan Ghai, Shamit Lal, Shivam Duggal, and Shrey Manik (2017)** | To recognize emotions in speech and classify them in different emotion output classes. | The components of the study include the emotional speech database, feature extraction, feature vector, and classifiers, which are used to recognize emotions in speech signals. | The system extracts feature from audio signals, such as energy and Mel Frequency Cepstral coefficients, and uses supervised learning algorithms to classify and recognize emotions in speech signals. | The system leverages perceptual features including Mel Frequency Cepstral coefficients (MFCC), energy of the speech signals, and the classification algorithms Support Vector Machine (SVM), Random Decision Forest, and Gradient Boosting. |  |  |  | The system achieved a maximum accuracy of 81.05% in recognizing emotions in speech signals, with the Random Decision Forest classifier providing the highest accuracy. | 1. High accuracy in recognizing emotions in speech signals.  2. Effective use of Mel Frequency Cepstral coefficients and energy of speech signals as feature inputs. | The system may misclassify happiness as anger and is susceptible to performance variations due to noise and emotional expression variability. |  | The proposed approach achieved high accuracy of 81.05% in recognizing emotions in speech signals, with Random Forest algorithm performing the best among the classifiers tested. |
| **Apeksha Aggarwal, Akshat Srivastava, Ajay Agarwal, Nidhi Chahal, Dilbag Singh, Abeer Ali Alnuaim, Aseel Alhadlaq, and Heung-No Lee (2022)** | To create and evaluate effective methods for Speech Emotion Recognition | The system comprises two-way feature extraction methods, Principal Component Analysis (PCA), Deep Neural Network (DNN) and pre-trained VGG-16 model. | The system utilizes two different methods of feature extraction to improve the effectiveness of emotion recognition. | Feature extraction plays a crucial role in the process of recognizing speech emotions. Utilizing PCA for feature extraction with DNN and utilizing pre-trained VGG-16 model for speech emotion recognition. |  |  |  | The performance of the proposed system is evaluated using multiple algorithms and two datasets. The RAVDESS dataset is found to provide significantly better accuracy than using numeric features on a DNN. | This model helps in dataset reduction, which may enhance model performance and generalization. | Limited Emotion classes, models are computationally expensive. |  | The proposed approach for the RAVDESS dataset, achieved an accuracy of 72%. For TESS dataset using VGG-16 model, it achieved accuracy of 90%. |
| **Tedd Kourkounakis, Amirhossein Hajavi, Ali Etemad(2020)** | The work goal of the paper is to address speech disfluencies and stutters in the workplace. | Feature extraction, recurrent layers, detection task, data, and annotation | FluentNet uses a combination of deep neural network techniques to detect speech disfluency. Additionally, FluentNet incorporates an attention mechanism to focus on the important parts of speech, allowing for better performance in detecting disfluencies. | This model incorporates a Squeeze-and-Excitation Residual Network (SE-ResNet) and bidirectional LSTM layers for effective stutter feature learning. |  |  |  | FluentNet achieves an average miss rate of 9.35% and an accuracy of 91.75%, surpassing other models and setting a new state-of-the-art. It outperforms previous models and benchmark models on the LibriStutter dataset as well. | The proposed model utilizes direct audio signals, spectrogram features and its innovative architecture providing an efficient framework for learning stutter specific features. | Limitations of the FluentNet includes challenges in classifying interjections, lack of sufficient training data and poor performance on word repetitions and prolongations. |  | On the UCLASS dataset, FluentNet achieves an average miss rate of 9.35% and an accuracy of 91.75%, surpassing other models and setting a new state-of-the-art. |
| **Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, and Rajesh Kumar Muthu (** | To identify speaker’s emotion. | The systems components include the input speech signal, feature extraction using MFCC and LPC, classification based on SVM, and the output. | The mechanism of the system involves the following steps:  1. Procuring the speech signal  2. Extracting features using MFCC and LPC  3. Using SVM to classify the features into four emotions: sadness, anger, fear, and happiness. | The features used in this system include MFCC (Mel-frequency cepstral coefficients), LPC (Linear Predictive Coding), pitch, energy, and speaker rate. These features are extracted from the input speech signal and used for emotion classification using the SVM algorithm. |  |  |  | The system achieved emotion classification using SVM with an accuracy of 85% on the UGA and LDC datasets, demonstrating its effectiveness in speech emotion recognition. | The advantages of the system include its ease of training, scalability to high-dimensional data, and the ability to handle non-linear classification tasks effectively. | The limitations may include the need to carefully select appropriate kernel functions for non-linear classification tasks and computational cost is high. |  | The system achieved decently good results in speech emotion recognition, with an accuracy of 85% using SVM on the UGA and LDC datasets for classification of emotions. |