**Literature Review (Secondary Research) Template**

|  |  |
| --- | --- |
| **Student Name** | **Behara Amulya** |
| **Project Topic Title** | **SpeechSentio – AI powered speech therapy with emotion analysis** |

|  |
| --- |
| **1** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| https://ieeexplore.ieee.org/document/9053893 | | [Tedd Kourkounakis](https://ieeexplore.ieee.org/author/37088481517)  [Amirhossein Hajavi](https://ieeexplore.ieee.org/author/37088482795)  [Ali Etemad](https://ieeexplore.ieee.org/author/37087028037) | | | Speech, stuttering, disfluency, deep learning, residual network, LSTM | |
| **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?** | |
| Deep residual network with bidirectional long short-term memory layers. | | The goal of this solution is to detect and classify different forms of stutters. The problem that needs to be solved is that there is minimal data and research on the identification and classification of stuttered speech. | | | bidirectional long short- term memory (Bi-LSTM), spectrogram feature vectors, batch normalization and ReLu activation functions, root means square propagation (RMSProp) optimizer, softmax loss function, Nvidia 1080 Ti GPU | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | |
| The paper's mechanism involves data collection, feature extraction, deep learning model training, and performance evaluation to detect and classify stutter disfluencies.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | 1 | Speech samples are collected from the University College London's Archive of Stuttered Speech (UCLASS) dataset. The dataset contains samples of monologues from participants with known stuttered speech impediments. Each recording is manually annotated for one of seven stutter disfluencies. | Using a real-world dataset with manual annotations provides a diverse and challenging dataset for training and evaluation. | The dataset may have limitations in terms of size and diversity, and manual annotation can be time-consuming. | | 2. | Spectrogram features are extracted from audio clips. Spectrograms represent audio signals in a format suitable for deep learning | Spectrograms are commonly used in speech analysis, and they capture the temporal and spectral characteristics of the speech signal. | Spectrograms may not capture all relevant information, and the choice of spectrogram parameters (e.g., window size) can impact results. | | 3. | A deep residual neural network is used for feature embedding. The network architecture includes convolutional layers with batch normalization and ReLU activation functions. | Deep residual networks are known for their ability to capture complex features. The architecture can learn stutter-specific features effectively. | Training deep networks can be computationally expensive and requires a large amount of data. | | 4. | The learned feature embeddings are processed by bidirectional LSTM layers. Each LSTM layer consists of 512 bidirectional LSTM units. | LSTM layers are effective in modeling sequential data. Bidirectional LSTMs capture context from both past and future embeddings. | The choice of hyperparameters (e.g., LSTM units, dropout rates) can affect model performance. | | 5. | The model is trained using TensorFlow's Keras API with a root mean square propagation (RMSProp) optimizer and softmax loss function. Leave-one-subject-out (LOSO) cross-validation is used for rigorous testing. | Cross-validation ensures robust evaluation. The use of bidirectional LSTMs and deep networks contributes to the model's performance. | Training deep learning models can be time-consuming and may require significant computational resources. | | 6. | The model's performance is evaluated in terms of accuracy and miss rate for each class of stutter disfluency. | The evaluation provides insights into the model's ability to detect and classify different stutter types. | Evaluating the model's performance can be challenging due to the need for diverse and well-annotated datasets. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | **Stutter Type Classification Outcome:**  The outcome variable involves the categorization of distinct stutter types, such as sound repetition, word repetition, phrase repetition, revision, interjection, or prolongation, determined through the proposed deep residual network and bidirectional long short-term memory (LSTM) model. | **Acoustic Spectrogram Feature Vectors and Model Architecture:**  The independent variables encompass acoustic features, specifically spectrogram feature vectors derived from audio clips. Additionally, the architectural components of the deep learning model, including the deep residual network and bidirectional LSTM layers, contribute to the independent variables. | **Dataset Magnitude and Severity of Stutter Impairment:**  The moderating variables involve the scale of the dataset, characterized by the number of participants and audio samples, and the severity of stutter in participants. These factors may modulate the relationship between independent and dependent variables. | **Learned Feature Embeddings and Residual Network:**  The mediating variables consist of the feature embeddings acquired through the deep residual network. These embeddings serve as an intermediary step in capturing stutter-specific characteristics, addressing challenges such as the vanishing gradient problem during model training. | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In this study, the dependent variable is the classification outcome for different stutter types, influenced by acoustic spectrogram feature vectors and a deep residual network with bidirectional long short-term memory layers. The dataset, with various stutter severity levels, moderates the classification effectiveness. The mediating variable, represented by learned feature embeddings, acts as a crucial link between input features and classification. The interplay among these variables highlights the importance of dataset characteristics, model architecture, and quality of learned features in achieving robust stutter type classification based solely on acoustic features. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Audio speech signals | Classification of different types of stutter disfluencies | | | | This solution integrates deep learning, residual networks, and bidirectional LSTMs to classify various stutter disfluencies from audio speech signals. It achieves state-of-the-art performance and holds potential for future research, including multi-class classification. Additional feature selection methods may further enhance results. | | | The contribution of this work lies in the development of a robust system for detecting and classifying various types of stutter disfluencies using deep learning, residual networks, and bidirectional LSTMs. This methodology offers a significant advancement in the field of speech disfluency analysis by avoiding the reliance on language models or ASR, making it more efficient and accurate. The system's potential applications in therapy, speech monitoring, and improving presentation skills hold the promise of positively impacting the lives of individuals with stuttering impediments. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| This solution holds the potential to greatly improve the diagnosis and therapy of stuttering, benefiting millions affected by this speech impediment. By offering robust and efficient stutter detection through acoustic features, it enhances early intervention and therapeutic success rates. Its application in speech analysis can contribute to a higher quality of life for individuals with stuttering disorders. | | | | While this solution offers valuable advancements, its reliance on acoustic features alone may lead to occasional misclassification, particularly for more complex stutter types. Additionally, it may not address gender, accent, and speech rate variations, limiting its effectiveness in diverse populations. The potential for false negatives in classification could impact the accuracy of stutter detection. | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work presents an innovative approach to stutter detection by focusing solely on acoustic features and avoiding ASR, enhancing efficiency. However, its performance limitations, such as occasional misclassification and sensitivity to variations, warrant further exploration for robust real-world applications. | | | TensorFlow, Keras, Nvidia CUDA Toolkit, Librosa, Scikit-learn, Jupyter Notebook, Audacity, MATLAB, PyTorch, GitHub | | | 1. Abstract 2. Introduction 3. Related Work 4. Proposed Methodology 5. Experiment Setup And Results 6. Conclusion and Future work 7. Acknowledgements |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Version 2.0 Week 2** | | | | | | |
| **2** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| https://www.sciencedirect.com/science/article/pii/S1877050920318512 | | Hadhami Aouani  Yassine Ben Ayed | | | Emotion recognition, MFCC, ZCR, TEO, HNR, SVM, auto-encoder | |
| **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?** | |
| Feature-based SVM Emotion Recognition with Auto-Encoder-Based Feature Dimension Reduction. | | The goal of this solution is to recognize human emotions in speech data using a combination of acoustic features (MFCC, ZCR, TEO, HNR) and machine learning techniques, primarily Support Vector Machines (SVM). The problem that this solution aims to address is the automatic recognition of human emotions in speech data. | | | Feature extraction method**(**MFCC, ZCR, TEO, HNR), Classification model (SVM with Linear, Polynomial & RBF kernel), Feature Dimension Reduction Techniques (Audo Encoder), Ryerson Multimedia Lab (RML) emotion database, Evaluation Metrics | |
| **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 acoustic features, including Mel-frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Teager Energy Operator (TEO), and Harmonic-to-Noise Ratio (HNR), are extracted from audio samples. | These features capture important information related to emotions in speech signals, and they are widely used in speech and emotion recognition. | The feature set can be high-dimensional and may contain redundant or irrelevant information. | | 2 | Support Vector Machines (SVM) are employed with different kernels (Linear, Polynomial, and RBF) to classify emotions based on the extracted features. | SVM is a robust classification algorithm, and different kernels allow for flexibility in modeling complex decision boundaries. | SVM performance can be sensitive to kernel and parameter selection. It may not perform optimally with high-dimensional feature sets. | | 3 | Auto-encoders, including Basic AE and Stacked AE, are used to reduce the dimensionality of the feature set by learning a compact representation. | Dimension reduction helps mitigate the curse of dimensionality, reduces computational complexity, and can lead to improved model performance. | The choice of auto-encoder architecture and hyperparameters can impact the quality of the reduced representation. AE training may require careful tuning. | | 4 | Parameters, such as the number of hidden units in auto-encoders, the number of iterations, weight regularization, and SVM kernel parameters, are fine-tuned. | Parameter optimization aims to maximize the system's accuracy and recognition rates. | Parameter tuning can be time-consuming and may overfit the model to the training data. | | 5 | The system's performance is evaluated using recognition rates and accuracy for each emotion class. | Metrics provide insights into the system's ability to recognize specific emotions accurately. | Evaluation metrics do not provide information about misclassification patterns or the system's generalization to unseen data. | | 6 | The proposed system is compared with other state-of-the-art methods in emotion recognition. | Comparative analysis helps identify the system's strengths and areas where it outperforms existing approaches. | Results may vary based on the choice of emotion databases and evaluation criteria. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | **Emotion Classification Outcome**:  The dependent variable in this study is the Emotion Classification Outcome, involving the categorization of emotions based on extracted features. The SVM and Auto-Encoder (AE) combined model aims to classify emotions using parameters such as 39 Mel Frequency Cepstral Coefficients (MFCC), Harmonic-to-Noise Ratio (HNR), Zero Crossing Rate (ZCR), and Teager Energy Operator (TEO). | **Acoustic and Prosodic Features:**  The independent variables consist of acoustic and prosodic features extracted from the audio signals. These features include 39 MFCC, HNR, ZCR, and TEO, which collectively form the feature vectors used in the classification algorithm. The SVM and AE act on these independent variables for emotion classification. | **Feature Dimension Reduction Method (Auto-Encoder):**  The moderating variable in this study is the Feature Dimension Reduction Method, specifically the use of Auto-Encoder (AE). AE serves as a moderating factor in the recognition of emotions by extracting relevant features from the initial parameters, contributing to improved SVM performance. | **Learned Feature Representation (Auto-Encoder Output):**  The mediating variable involves the Learned Feature Representation obtained through the output of the Auto-Encoder. The AE model learns a reduced informative representation of the data, acting as an intermediary step in feature selection before inputting the refined features into the SVM classifier. | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The Independent Variable (IV) encompasses acoustic and prosodic features like 39 MFCC, HNR, ZCR, and TEO. These features directly impact the Dependent Variable (DV), representing the emotion classification outcome. The Feature Dimension Reduction Method, serving as a Moderating Variable, employs an Auto-Encoder to compress input features before SVM classification. Simultaneously, the Auto-Encoder's learned feature embeddings act as the Mediating Variable, optimizing the relationship between the Independent and Dependent Variables. This intricate interplay enhances the effectiveness of emotion recognition in the proposed system. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Audio samples containing human speech -RML emotion database | Classification and recognition of emotions in audiovisual data. | | | | This solution comprises audiovisual emotion recognition using fused audio features like MFCC, HNR, ZCR, and TEO, enhanced with auto-encoders for feature reduction. SVM is employed for emotion classification. It surpasses existing techniques, particularly in emotion identification from audiovisual data. | | | This work contributes an effective emotion recognition system by combining HNR with traditional audio features and utilizing auto-encoders for feature reduction, yielding improved accuracy. It holds potential for broader applications and larger datasets. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| This solution enhances the accuracy of emotion recognition in audio data, which is valuable for various applications like human-computer interaction and sentiment analysis in speech. It provides a foundation for more reliable emotion detection in diverse linguistic contexts, improving user experiences and data-driven decision-making. | | | | One potential negative impact could be the increased computational complexity due to feature extraction and dimension reduction, which might not be suitable for real-time applications with limited resources. | | |
| **Analyse This Work by Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work introduces an innovative approach to emotion recognition using a combination of acoustic features and auto-encoders for dimension reduction, showing promising results. However, it may lack generalizability to different datasets and requires further evaluation across diverse linguistic and emotional contexts. | | | Scikit-learn, TensorFlow, PyTorch, Matplotlib, Seaborn | | | 1. Abstract 2. Introduction 3. Related work 4. Methods 5. Proposed system 6. Experiments and results 7. Conclusion |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

**--End of Paper 2—**

|  |
| --- |
| **3** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| https://link.springer.com/article/10.1007/s10772-018-09572-8 | | S. Lalitha  Shikha Tripathi  Deepa Gupta | | | Arousal, BFCC, Cepstrum , DNN , Emotion detection , Perceptual features , Recognition accuracy , Valence | |
| **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?** | |
| Deep neural networks (DNN) along with various perceptual speech features for speech emotion recognition (SER) | | The goal is to develop a Speech Emotion Recognition (SER) system using DNNs and perceptual features to classify and map human emotions in spoken language. The problem is recognizing emotions in speech, both categorically and continuously, using the Berlin corpus as the dataset. | | | Pre-processing module(Involves framing and windowing without filtering the speech signal), Speech feature extraction module(Extracts various speech features from Mel, Bark, and Inverted Mel filter banks, as well as additional features like fundamental frequency, LPC, and functionals), Classification module(DNN), Performance metrics | |
| **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 | Pre-processing includes framing and windowing without any filtering of the speech signal. It prepares the input data for feature extraction. | Keeps the speech signal intact without filtering, which may be important for preserving emotional information. | May not address noise or other signal quality issues. | | 2 | In Speech feature extraction various speech features are extracted from Mel, Bark, and Inverted Mel filter banks, as well as additional features like fundamental frequency, LPC, and functionals. | Provides a rich set of features that capture different aspects of the speech signal related to emotions. | Increases the dimensionality of the feature space, potentially leading to increased computational complexity. | | 3 | A Feed-Forward Back Propagation Network (DNN) is used for emotion classification based on the extracted features. | DNNs are capable of modeling complex, nonlinear relationships in the data, which can improve emotion classification performance. | Training DNNs can be computationally intensive, and they may require a large amount of labeled data for effective training. | | 4 | The system's performance is evaluated using recognition accuracy and confusion matrix in both categorical and continuous emotion dimensions. | Provides quantitative measures of the system's ability to recognize emotions. | Performance evaluation metrics do not provide insights into the interpretability of the DNN model or its generalization to other datasets. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | **Emotion Recognition Performance**: This variable refers to the system's effectiveness in recognizing emotions from speech signals.  **Examples:** Recognition accuracy (%) in both 1-dimensional (categorical) and 2-dimensional (valence and arousal) emotion spaces. | **Perceptual Features (MFCC's, PLPC, MFPLPC, BFCC, RPLP, IMFCC):** Auditory cues derived from speech signals, including Mel frequency cepstral coefficients (MFCC's), perceptual linear predictive cepstrum (PLPC), Mel frequency perceptual linear prediction cepstrum (MFPLPC), bark frequency cepstral coefficients (BFCC), revised perceptual linear prediction coefficient’s (RPLP), and inverted Mel frequency cepstral coefficients (IMFCC). | Potential moderating factors could include variations in speech patterns among different individuals or environmental conditions during recording. | **Perceptual Features (MFCC, PLPC, MFPLPC, BFCC, RPLP, IMFCC):** Speech features acting as mediators between the independent variable (input to the system) and the dependent variable (emotion recognition performance). | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The relationship among these variables in the article involves the interplay between the chosen perceptual features, the architecture of the DNN, and their combined impact on the system's ability to recognize emotions from speech signals. The absence of an explicitly mentioned moderating variable suggests a focus on the direct influence of the chosen variables on emotion recognition performance. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Audio signal containing human speech, specifically speech samples from the Berlin corpus | Prediction or classification of the emotional state conveyed in the input audio signal | | | | This solution combines deep neural networks, such as Feed-Forward Back Propagation Networks, with perceptual speech features derived from Mel and Bark filter banks. It enables the classification of various emotions in speech, including fear, anger, boredom, and more, in both categorical and continuous emotion spaces. The system offers a compact feature vector, speaker independence, and high recognition accuracy, demonstrating its effectiveness for emotion detection in audio signals. The potential for multi-corpus and multimodal applications, as well as insights into addressing imbalanced datasets, adds to its feature set. | | | This work contributes by systematically exploring and identifying significant perceptual features for emotion detection in speech, enabling improved recognition accuracy. The utilization of Feed-Forward Back Propagation Networks offers an effective classification mechanism, particularly when combined with the selected features. It adds value by enhancing the understanding of emotional content in speech, applicable to various domains such as human-computer interaction and sentiment analysis. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| In the context of this project, the positive impact of this solution lies in its ability to enhance emotion recognition accuracy in speech, which is crucial for applications like human-computer interaction, sentiment analysis, and affective computing. The identified perceptual features and DNN classification method offer a valuable tool for understanding and extracting emotional content from audio data, ultimately improving the project's performance and usability in emotion-related tasks. | | | | It relies on significant computational resources and specific datasets, limiting its scalability and generalizability to different languages and cultures. Additionally, the feature engineering process may require domain expertise, hindering accessibility for non-experts. | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work offers a promising approach to speech emotion recognition (SER) by combining deep neural networks (DNNs) with selected perceptual speech features. However, while it demonstrates competitive performance, there is room for improvement in feature selection and architecture choice to enhance recognition across diverse datasets and real-world applications. | | | TensorFlow, PyTorch, Keras, NumPy, pandas, Matplotlib, Scikit-learn | | | 1. Abstract 2. Introduction 3. Related work 4. Data source 5. Proposed system 6. Experimental set‑up 7. Experimental results and analysis 8. Performance evaluation 9. Conclusion and outlook |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

**--End of Paper 3--**

|  |
| --- |
| **4** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| https://ieeexplore.ieee.org/document/9330527 | | [Bassam Ali Al-Qatab](https://ieeexplore.ieee.org/author/38276950600)  [Mumtaz Begum Mustafa](https://ieeexplore.ieee.org/author/38241272200) | | | Acoustic features, automatic dysarthric speech recognition system, dysarthria, classification algorithms, feature selection methods. | |
| **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?** | |
| Automatic speech recognition (ASR) systems | | The goal of this solution is to improve the accuracy of automatic speech recognition (ASR) systems for dysarthric speakers. The problem is that ASR systems are not very accurate when used with dysarthric speakers. | | | Feature selection methods: Interaction Capping (ICAP), Conditional Information Feature Extraction (CIFE), Conditional Mutual Information Maximization (CMIM), Double Input Symmetrical Relevance (DISR), Joint Mutual Information (JMI), Conditional redundancy (Condred) and Relief.  Classification algorithms: Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Classification and Regression Tree (CART), Naive Bayes (NB), and Random Forest (RF).  Acoustic features : prosody, spectral, cepstral, and voice quality. | |
| **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 | selecting the NEMOURS database, which contains recorded speech from a single dysarthric speaker with varying levels of severity. | The NEMOURS database offers a standardized and consistent dataset for analysis, ensuring the replicability of research. | The limited number of speakers may not fully represent the diversity of dysarthric speech. | | 2 | A wide range of acoustic features is extracted from the speech data, including prosodic, spectral, cepstral, and voice quality features. | Extracting multiple features enables a comprehensive analysis of different aspects of dysarthric speech, potentially capturing valuable information. | The large number of features can make subsequent classification challenging due to dimensionality issues, requiring feature selection. | | 3 | The study employs a feature selection process to reduce the dimensionality of the data. A method based on a formula is used to select a specific number of features. | Feature selection reduces computational complexity, and it can help identify the most relevant features, potentially improving classification accuracy. | The choice of the feature selection method may not be definitive, and the formula used may not be the best fit for all datasets. Different methods could yield different results. | | 4 | Six different classification algorithms, including Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Classification and Regression Tree (CART), Naive Bayes (NB), and Random Forest (RF), are used to classify the severity of dysarthric speech. | The variety of classification algorithms allows for a comprehensive evaluation, identifying which approach works best for the problem. | The choice of classification algorithms can significantly impact the results, and it may be challenging to determine the best approach. | | 5 | The classification models' performance is evaluated using k-fold cross-validation with k=10, providing a robust estimate of accuracy. | Cross-validation reduces overfitting risks and offers a more reliable assessment of classifier performance. | The choice of the number of folds (k) can affect the results, and the study focused on accuracy as the primary evaluation measure, which may not capture all aspects of performance. | | 6 | Friedman's m statistics are applied to rank the classifiers and feature selection methods based on their classification accuracy. | Ranking helps identify the most effective combinations of classifiers and feature selection methods, simplifying the selection of the best approach. | Rankings are based on accuracy measures, and any bias in these measures could lead to incorrect rankings. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | **Target Variable - Dysarthric Speech Severity Classification:** This represents the primary outcome variable, categorized into severity levels, serving as the target for the classification. | a. Acoustic Features:   * Prosody, Spectral, Cepstral, and Voice Quality Parameters   b. Feature Selection Methods:   * Interaction Capping (ICAP), Conditional Information Feature Extraction (CIFE), Conditional Mutual Information Maximization (CMIM), Double Input Symmetrical Relevance (DISR), Joint Mutual Information (JMI), Conditional Redundancy (Condred), and Relief | **Speech Database Size:** This variable moderates the relationship between the independent variables (acoustic features, feature selection methods) and the dependent variable by influencing the effectiveness of the ASR system. | **Feature Selection Methods:** Specifically, the algorithms like ICAP, CIFE, CMIM, DISR, JMI, Condred, and Relief act as mediating variables. They intervene in the relationship between the independent variables (acoustic features) and the dependent variable (Dysarthric Speech Severity Classification). | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The acoustic features play a direct role in determining the Dysarthric Speech Severity Classification. However, the impact of these features is shaped and mediated by the chosen feature selection methods. The moderating variable, speech database size, further influences the overall relationship by adjusting the effectiveness of the ASR system in classifying dysarthric speech. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Dysarthric speech | Methodology for improving ASR | | | | This is simply an integration of four acoustic features and seven feature selection methods to design a hybrid one. We can still integrate other feature selection methods which gives us even more better results. | | | Designing hybrid classifier is a good thought, where four acoustic features and seven feature selection methods working together to resolve individual issues. The proposed methodology is a valuable contribution to the field of dysarthric speech recognition. The hybrid classifier has the potential to improve the quality of life for people with dysarthria by improving the accuracy of ASR systems. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| Hybrid classifier has the potential to improve the accuracy of ASR systems for dysarthric speakers by integrating individual advantages of acoustic features and feature selection methods.Overall, the proposed methodology is a valuable contribution to the field of dysarthric speech recognition. | | | | The proposed methodology is not computationally efficient and it is trained on a specific dataset that may not be generalizable to other populations of dysarthric speakers. | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The proposed methodology is a valuable contribution to the field of dysarthric speech recognition. The authors make a valuable contribution by integrating existing algorithms in a new way to improve the accuracy of ASR systems for dysarthric speakers. | | | openSMILE, MATLAB, Libsvm, Random Forest(RF) | | | 1. Abstract 2. Introduction 3. Related Work 4. Proposed Methodology 5. Experiments and Results 6. Conclusion |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

**--End of Paper 4—**

|  |
| --- |
| **5** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| https://www.researchgate.net/publication/321243350\_Detecting\_Stuttering\_Events\_in\_Transcripts\_of\_Children's\_Speech | | Sadeen Alharbi  Madina Hasan  Anthony J H Simons  Shelagh Brumfitt &  Phil Green | | | Stuttering event detection, Speech disorder, human-computer interaction, CRF, HELM | |
| **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?** | |
| Machine learning approaches, specifically the "Hidden Event Language Model (HELM)" and "Conditional Random Fields (CRF). | | The goal of the solution is to automatically detect and classify stuttering events in transcripts of children's speech. This addresses the problem of accurately identifying stuttering in speech, which can impact a child's development if left untreated. The solution aims to provide a faster and more reliable method for early diagnosis and intervention. | | | Data Transcription and Annotation, Data Normalization and Feature Extraction ( extracting word-level features, including n-grams and post-words), Classification Approaches (HELM & CRF), Data Augmentation, Metrics( precision, recall, F1 score, and accuracy). | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | |
| This process combines data collection, preprocessing, machine learning models, data augmentation, and evaluation to detect and classify stuttering events in children's speech transcripts.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | 1 | The work starts by collecting orthographic transcripts of children's speech, which serves as the input data. Annotators label different types of stuttering events within the text using a specific annotation approach. | This step provides the necessary data to train and test the classification models, making it possible to analyze stuttering in children's speech. | Collecting and annotating the data can be time-consuming and may rely on the availability of such data. | | 2 | Text normalization is performed to prepare the data for analysis. This step includes transforming text entities (e.g., dates, numbers, and times) into words and extracting word-level features, including n-grams and post-words. | Data preprocessing ensures that the data is in a suitable format for machine learning analysis, making it easier to extract relevant features. | The choice of features and the preprocessing process may impact the quality of the data and the results. | | 3 | The work employs two machine learning approaches, HELM and CRF, to classify and detect stuttering events within the transcripts based on the extracted features. | Machine learning models can automatically detect stuttering events in transcripts, potentially reducing the subjectivity associated with manual assessment. | The performance of the models depends on the quality and quantity of training data and the choice of features. | | 4 | To improve the models' ability to detect stuttering events, the work generates additional sentences with stuttering patterns, supplementing the original training data. | Data augmentation increases the diversity and volume of the training data, enhancing the models' performance, especially for rare stuttering events. | The quality of augmented data may vary, and it could introduce noise into the training data. | | 5 | The solution employs standard evaluation metrics such as precision, recall, F1 score, and accuracy to assess the performance of the models in detecting stuttering events in the transcripts. | Evaluation metrics provide a quantitative measure of the solution's effectiveness in identifying stuttering, enabling comparisons and improvements. | The choice of evaluation metrics may not capture all aspects of model performance, and some stuttering events may remain challenging to detect. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| <Find all main factors and variables that are related to each solution. 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** | | Stuttering Event Detection Accuracy | Machine learning approaches (HELM and CRF).  Data augmentation techniques. | Training Data Availability. | Text Normalization features  Word-level features  In the context of the classification approaches:  **1.HELM (Hidden Event Language Model):**  Probability of stuttering events at the end of each observed word given its context.  **2.CRFs (Conditional Random Fields):**  Sequence labelling and segmentation of stuttering events.  Estimation and optimization of the posterior probability of the label sequence given a sequence of features. | | | | | | | |
| |  | | --- | | **Relationship Among the Above 4 Variables in This article** | | The chosen machine learning approaches and data augmentation techniques influence the accuracy of detecting stuttering events. However, the presence or absence of sufficient training data moderates this relationship. The process of text normalization and the extracted word-level features serves as a mediating mechanism, influencing how the machine learning approaches impact the accuracy of stuttering event detection. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | orthographic transcripts of children's speech. | Diagnosis of stuttering events in children's speech transcripts. | | | | The solution's key feature is its use of machine learning approaches (HELM and CRF) to detect and classify stuttering events in children's speech transcripts, along with data augmentation to improve performance for rare events. | | | The work contributes to the field of stuttering event detection by introducing machine learning techniques and data augmentation to enhance accuracy and addresses the scarcity of training data. Its value lies in improving the assessment of stuttering in children, aiding early intervention, and expanding the annotated speech data available for research. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The positive impact of this solution in the project domain is more accurate and efficient stuttering event detection in children's speech, facilitating early intervention and research in speech disorders. | | | | The negative impact of this solution in the project domain might be the reliance on generated data for training, which could introduce noise and inaccuracies in the models, potentially affecting the precision of stuttering event detection. | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work leverages machine learning models like HELM and CRF for automated stuttering event detection in children's speech transcripts, potentially aiding early intervention in speech disorders. However, challenges persist in detecting less frequent stuttering events and ethical aspects related to data augmentation. | | | Hidden Event Language Model (HELM), Conditional Random Fields (CRF), Automatic Speech Recognition (ASR), and the SRILM toolkit for data augmentation | | | 1. Abstract 2. Introduction 3. Data Transcription and Annotation 4. Data Normalisation and Features Extraction 5. Classification Approaches 6. Data Augmentation 7. Metrics 8. Experiments 9. Conclusions and Future Work |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

**--End of Paper 5--**

**Work Evaluation Table**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Work Goal** | **System's Components** | **System's Mechanism** | **Features /Characteristics** | **Cost** | **Speed** | **Security** | **Performance** | **Advantages** | **Limitations /Disadvantages** | **Platform** | **Results** |
| **Hadhami Aouani , Yassine Ben Ayed(2020)** | To create a speech emotion recognition system. | The system consists of two stages: feature extraction and classification engine. | This is an article about speech emotion recognition. It discusses using MFCC features and an autoencoder to improve the accuracy of speech emotion recognition. | It discusses two sets of features, MFCC and Teager Energy Operator. The Support Vector Machines (SVM) is used as a classifier method. | $1,000 and $10,000 for a basic | SER with MFCCs is fast |  | Proposed system with SVM Stacked AE 39MFCC, ZCR, TEO, HNR filtered by stacked AE shows accuracy of 74.07 | 1.Comprehensive Feature Utilization  2.Optimized Classification with Auto-Encoder | 1.Dependency on Feature Relevance  2.Sensitivity to Hyperparameter Tuning  3.Generalization Challenges  4.Complexity of Auto-Encoder Training  5.Limited Interpretability |  | The proposed emotion recognition system achieved recognition rates of 72.83% and 74.07% using a basic auto-encoder and stacked auto-encoder for feature dimension reduction, surpassing the system without feature selection (65.43%) on the RML emotion database. |
| **S. Lalitha, Shikha Tripathi and Deepa Gupta(2018)** | The paper aims to explore the effectiveness of perceptual-based speech featuresfor emotion detection using deep neural networks on the Berlin database. | The system comprises pre-processing, feature extraction (utilizing Mel, Bark, and inverted Mel filter banks, as well as additional features), and classification using a deep neural network | The system involves preprocessing audio data, extracting perceptual features (MFCCs, PLPC, BFCC, RPLP, MFPLPC, IMFCC), and utilizing a deep neural network for effective emotion detection, achieving improved performance on the Berlin database. | The system leverages perceptual features including MFCCs, PLPC, BFCC, RPLP, MFPLPC, and IMFCC for enhanced emotion detection, optimizing recognition accuracy on the Berlin database. |  |  |  | The performance of the system, evaluated using deep neural networks with perceptual speech features, shows improved emotion recognition accuracy compared to conventional methods, as demonstrated on the Berlin database. | The advantages of the proposed system include enhanced speech emotion detection accuracy through the utilization of perceptual-based features, such as Mel frequency cepstral coefficients (MFCCs), with DNNs | The limitations of the system may include potential challenges in handling diverse emotional expressions, dependency on the selected features, and sensitivity to variations in the training dataset, affecting generalization to real-world scenarios. |  | The results of the system indicate improved emotion recognition accuracy, particularly in valence and arousal dimensions, using a combination of perceptual-based speech features and deep neural networks, as demonstrated on the Berlin emotion database. |
| **Sadeen Alharbi, Madina Hasan, Anthony Simons, and Shelagh Brumfitt(2017)** | To investigate the feasibility of using machine learning for stuttering detection in children's speech transcripts. | The system comprises three main components: pre-processing, speech feature extraction, and classification modules, utilizing machine learning approaches such as HELM and CRF for stuttering event detection in children's speech transcripts. | The system employs machine learning approaches, specifically HELM and CRF, to detect stuttering events in children's speech transcripts, addressing challenges such as limited training data and high dimensionality. | The system utilizes features such as n-grams (up to 4-grams) and post-words extracted from normalized transcripts for detecting stuttering events, considering various stuttering types (sound repetitions, part-word repetitions, word repetitions, phrase repetitions, prolongations). |  |  |  | The performance of the proposed system is evaluated through machine learning approaches, HELM and CRF, for detecting stuttering events in transcripts of children's speech, achieving improved results with CRF and data augmentation. | The advantages of the system include the application of machine learning techniques (HELM and CRF), successful detection of stuttering events in children's speech transcripts, and enhanced performance with data augmentation, providing valuable insights for automated diagnosis. | Limitations include challenges related to the lack of training data and the high dimensionality of the data, with a focus on stuttering event detection in children's speech transcripts is highlighted. |  | Results indicate that CRF outperforms HELM by 2.2% in baseline experiments for stuttering event detection in children's speech transcripts. Data augmentation proves beneficial, particularly for rarely available events, contributing to improved system performance. |
| **Bassam Ali Al-Qatab and Mumtaz Begum Mustafa (2021)** | The work goal of this paper is to evaluate the effectiveness of acoustic features and feature selection methods in classifying dysarthric speech based on the severity of impairment, employing various classification algorithms, with a focus on assessing classification accuracy and ranking performance. | The system's components encompass acoustic features (prosody, spectral, cepstral, voice quality), feature selection methods (Interaction Capping, Conditional Information Feature Extraction, Conditional Mutual Information Maximization, Double Input Symmetrical Relevance, Joint Mutual Information, Conditional Redundancy, Relief), and classification algorithms (Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, Classification and Regression Tree, Naive Bayes, RF | The system operates by extracting acoustic features (prosody, spectral, cepstral, voice quality) from dysarthric speech, employing feature selection methods (Interaction Capping, Conditional Information Feature Extraction, Conditional Mutual Information Maximization, Double Input Symmetrical Relevance, Joint Mutual Information, Conditional Redundancy, Relief), and utilizing classification algorithms (Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, CART, Naive Bayes, Random Forest) to categorize speech severity levels. | The study investigates dysarthric speech features, encompassing prosody, spectral, cepstral, and voice quality, using seven feature selection methods (Interaction Capping, Conditional Information Feature Extraction, Conditional Mutual Information Maximization, Double Input Symmetrical Relevance, Joint Mutual Information, Conditional Redundancy, Relief), and evaluates classification algorithms (Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, Classification and Regression Tree, Naive Bayes, Random Forest) for severity level classification. |  |  |  | The classification accuracy of the dysarthric speech analysis ranges from 40.41% to 95.80%, utilizing acoustic features and feature selection methods with six classification algorithms, including Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, Classification and Regression Tree, Naive Bayes, and Random Forest. | Advantages include its potential as assistive technology for individuals with speech impairments, leveraging acoustic features and diverse feature selection methods, achieving classification accuracy ranging from 40.41% to 95.80% with various classification algorithms. | Limitations of the dysarthric speech classification system include potential challenges associated with data sparsity, both in language coverage and speech database size, impacting the effectiveness of the automatic speech recognition (ASR) system. |  | The dysarthric speech classification system achieved classification accuracy ranging from 40.41% to 95.80%, utilizing acoustic features and various feature selection methods with six classification algorithms. |
| **Tedd Kourkounakis, Amirhossein Hajavi, Ali Etemad(2019)** | The goal of this work is to address the identification and classification of various forms of stuttering, focusing on acoustic features rather than language models. | The system's components include a deep RNN for learning stutter-specific features, bidirectional LSTM layers for sequential data analysis, and a classification module for identifying different types of stuttering based on acoustic features. | The system's mechanism involves generating spectrogram feature vectors from audio clips, utilizing a deep residual neural network to extract stutter-specific features, employing bidirectional long short-term memory (LSTM) layers for sequential data analysis | The system employs spectrogram feature vectors, a deep residual neural network (ResNet) with convolutional blocks, and bidirectional LSTM layers with dropout mechanisms for effective identification and classification of various stutter disfluencies based solely on acoustic features. |  |  |  | The proposed system achieves an average miss rate of 10.03% in detecting and classifying different forms of stutter disfluencies, surpassing the state-of-the-art by almost 27%. | The system offers advantages such as robust detection and classification of various stutter disfluencies solely based on acoustic features, outperforming existing models with an average miss rate of 10.03%. | The limitations of the system include potential misclassification of longer utterances due to reliance on four-second windows, and challenges in accurately identifying certain stutter types, such as prolongation, leading to a slightly higher miss rate for these cases. |  | The proposed model achieved an average miss rate of 10.03%, outperforming the state-of-the-art by almost 27%, in the detection and classification of different forms of stutter disfluencies using acoustic features. |