**Literature Review (Secondary Research) Template**

|  |  |
| --- | --- |
| **Student Name** | **K.Abhinay** |
| **Project Topic Title** | SpeechSentio: 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**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Version 1.0 \_ Week 1** | | | | | | | |
| **1** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| https://arxiv.org/abs/2105.05599 | | md.sahidullah | | | | Speech Recognition, Self-Supervised Learning ,Speech Representations, Automatic Speech Recognition (ASR), Language Modeling | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The goal of the research paper "Self-Supervised Speech Recognition with Speech Representations from ASR" is to improve the performance of automatic speech recognition (ASR) systems using self-supervised learning.** | | | | **Data preparation, Self-supervised learning objective, ASR training procedure** | |
| StutterNet: Stuttering Detection Using Time Delay Neural Network | | The aim of the paper is to propose a self-supervised learning approach for training speech representations. Speech representations are features that can be used to represent speech signals. | | | | The authors evaluate their proposed approach on a number of different ASR datasets, and show that it consistently outperforms other state-of-the-art self-supervised learning approaches. | |
| **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 authors use a variety of techniques to prepare the data for self-supervised learning, including data augmentation, noise injection, and masking. | The proposed approach has been shown to improve the performance of ASR systems, even when the ASR systems are trained on a small amount of labeled data. | The proposed approach is sensitive to the choice of hyperparameters, such as the learning rate and the batch size. | | **2** | The authors propose a self-supervised learning objective that is based on contrastive learning. In contrastive learning, the model learns to distinguish between positive and negative pairs of data. | The proposed approach reduces the amount of labeled data required to train ASR systems. This is important because labeled data is expensive and time-consuming to collect. | The proposed approach is prone to overfitting, especially when trained on a small amount of data. | | | | | | | | |
| **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** | | The primary focus of the research is on the accuracy of stuttering detection, which serves as the dependent variable. | StutterNet System Implementation including the use of TDNN(Time Delay Neural Network), acoustic features like MFCCs, and the specific architecture design. | factors influencing the relationship between the independent variable and the dependent variable | The optimization variables act as potential mediating variables, influencing the relationship between the StutterNet system implementation (independent variable) and stuttering detection accuracy (dependent variable). | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** |   **In this research, the StutterNet system's implementation (independent variable) directly influences the accuracy of stuttering detection (dependent variable). Optimization variables, such as layer size, context window, and filter bank size, serve as potential mediators, affecting the relationship between system implementation and detection accuracy.** | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | The input to the solution is unlabeled speech data. This data can be in any language and can be noisy. | The output of the solution is a set of speech representations. | | | | The key feature of the solution is that it uses self-supervised learning to train speech representations. This means that the model learns from unlabeled data, without the need for human-transcribed transcripts. | | | | The main contribution of the paper is the proposal of a new self-supervised learning objective for ASR. This objective is based on contrastive learning, where the model learns to distinguish between positive and negative pairs of data. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The proposed solution has been shown to improve the performance of ASR systems, even when the ASR systems are trained on a small amount of labeled data. | | | | | **Since this is a performance evaluation of various algorithms, not much to project on negative side as all the things used are defined in advance.** | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The research paper "Self-Supervised Speech Recognition with Speech Representations from ASR" proposes a new self-supervised learning objective for ASR. | | | | self-supervised learning | | | Abstract   1. INTRODUCTION 2. RELATED WORK 3. PROPOSED ARCHITECTURE 4. EXPERIMENTAL EVALUATION 5. RESULTS 6. 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://arxiv.org/abs/2302.09044 | | Colin Lea  colin\_lea@apple.com | | | | speech input, accessibility, stuttering, voice assistants, dictation | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) demonstrate how many common errors can be prevented, resulting in a system that cuts utterances of 79.1% less often and improves word error rate from 25.4% to 9.9%.** | | | | **Propose technical solutions to improve the performance of speech recognition systems for people who stutter.** | |
| Hybrid Decision Tree and Logistic Regression Classifier for email spam detection. | | Logistic Regression and decision tree - Stuter detection | | | | Handling data with noise | |
| **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 authors conducted two surveys to understand the needs and experiences of people who stutter when using speech recognition systems. | The authors' proposed solutions resulted in a 79.1% reduction in the number of times utterances were cut off and a 15.5% improvement in word error rate. | The authors' work is still in the early stages of development, and more research is needed to validate their findings and develop robust and scalable solutions. | | **2** | They also conducted technical investigations to identify and address common errors in speech recognition systems. | The authors' work could significantly improve the performance of speech recognition systems for people who stutter, making it easier for them to communicate and participate in everyday activities. | The authors' work is also limited to a specific type of speech recognition system, and it is not clear how well their solutions would generalize to other types of systems. | | **3** | The authors then proposed three technical solutions to improve the performance of speech recognition systems for people who stutter. |  |  | | **4** | The authors evaluated their proposed solutions on a real-world dataset of speech from people who stutter. |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The dependent variable is the Word Error Rate (WER), which is the key metric used to measure the accuracy of speech recognition. | Independent variables is the threshold set for endpoint detection in speech recognition. | moderating variable is Stuttering severity may moderate the impact of interventions. | A mediating variable is Annotations highlighting dysfluency types in the speech. | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In the context of speech recognition for people who stutter, the study investigates the impact of independent variables (Endpointer Threshold, ASR Decoder Tuning, Dysfluency Refinement) on dependent variables (Word Error Rate, Intent Error Rate), with stuttering severity serving as a moderating factor and dysfluency annotations as a mediating variable. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | The input to the system is a speech signal from a person who stutters. | The output of the system is a transcript of the speech signal. | | | | The key feature of the system is that it is robust to dysfluencies, such as repetitions and prolongations. | | | | The system can help people who stutter to communicate more effectively using speech recognition systems. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The proposed solution can make speech recognition systems more accessible to people who stutter. This is because the solution is robust to dysfluencies, such as repetitions and prolongations, which are common in stuttering speech. | | | | | **The proposed solution may reduce the accuracy of speech recognition systems in non-stuttering speech. This is because the solution is tuned to detect and correct dysfluencies, which are not present in non-stuttering speech.** | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| It involves assessing the innovative approaches used to enhance speech recognition for people who stutter. | | | | The tools employed for evaluation include dysfluency annotations, ASR models with varied architectures, and statistical methods such as Wilcoxon signed rank tests. | | | Abstract   1. Introduction 2. Related Work 3. Proposed Method 4. Experiment Results 5. 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://arxiv.org/abs/2106.06598 | | Shinji Watanabe | | | | speech sentiment analysis, pre-trained language model, pseudo label-based semi-supervised training speech sentiment analysis, pre-trained language model, pseudo label-based semi-supervised training | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The objective of the research paper is to propose a novel approach for speech sentiment analysis using pre-trained language models and a pseudo label-based semi-supervised training strategy.** | | | | **The paper propose a pseudo label-based semi-supervised training strategy to reduce the need for human-labeled data.** | |
| A Proposed Model for Leveraging Pre-trained Language Model for Speech Sentiment Analysis | | An approach to transfer knowledge from the written text to spoken text or speech domain using an LM | | | | MFCCs are a type of audio feature that is commonly used in speech processing. MFCCs are extracted from the speech signal and fed into the pre-trained language model. | |
| **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** | 2-step pipeline approach employs Automatic Speech Recognition (ASR) and transcripts-based sentiment analysis separately. | They can be used to train a sentiment analysis system on a large amount of data without human sentiment annotation. | They require a pre-trained language model, which can be expensive to train and deploy. | | **2** | Pseudo label-based semi-supervised training strategy uses a language model on an end-to-end speech sentiment approach. | They can be used to train a sentiment analysis system that is robust to noise in the speech signal. | The proposed approaches may not be as effective for speech sentiment analysis in low-resource languages or for speech sentiment analysis of specific domains, such as medical or legal speech. | |  |  |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | recall, precision, and F1 scores. | Text Source: ASR transcripts as input data | Quantity of Labeled Data: Labeled data acts as a moderating variable in the relationship between the independent variables and model performance. | Pseudo Label-based Semi-Supervised Training: Pseudo label-based semi-supervised training as a mediating variable between the independent variables and the dependent variable. | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In contemporary sentiment analysis research, the literature underscores the pivotal role of pre-trained language models, notably BERT, as an independent variable. These models significantly enhance the performance of speech sentiment analysis, representing a key focus in the exploration of cutting-edge methodologies for efficient sentiment classification in spoken language datasets. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Speech signal | Sentiment | | | | The system uses a pseudo label-based semi-supervised training strategy to reduce the need for human-labeled data. | | | | The system can be used to develop customer service chatbots that can understand and respond to customer queries in a more effective manner. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The proposed solution can be used to improve the performance of speech-based applications, such as customer service chatbots and virtual assistants. This is because the system can extract features from the speech signal that are informative for sentiment analysis. | | | | | **The proposed solution may reduce the accuracy of speech-based applications in certain cases. For example, the system may not be as effective for speech sentiment analysis in low-resource languages or for speech sentiment analysis of specific domains, such as medical or legal speech.** | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The research paper is well-written and the proposed approaches are well-motivated. The authors evaluate their approaches on a real-world dataset and show that they can improve F1 scores consistently compared to systems without a language model. The authors also acknowledge the limitations of their work and suggest directions for future research. | | | | The paper used these tools to evaluate their proposed approaches on a real-world dataset of speech. The results showed that the proposed approaches improved F1 scores consistently compared to systems without a language model. The confusion matrices showed that the proposed approaches were able to reduce the number of misclassified instances. | | | Abstract   1. Introduction 2. Related work 3. Approaches 4. Experiments 5. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

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

|  |
| --- |
| **4** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| https://www.mdpi.com/1424-8220/22/17/6369 | | Bagus Tris Atmaja and Akira Sasou | | | | affective computing; sentiment analysis; speech emotion recognition; sentiment analysis and emotion recognition; universal speech representation | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The objective of the research paper is to conduct a comprehensive review of the state-of-the-art in sentiment analysis and emotion recognition from speech using universal speech representations (USRs). USRs are a type of speech representation that is designed to be universal, meaning that they can be used for a variety of speech processing tasks, including sentiment analysis and emotion recognition.** | | | | **What are the components of it?** | |
| Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations | | Although in this study we conducted sentiment analysis and categorical emotion recognition independently, future studies could merge these tasks into a multitask learning approach, predicting both sentiment and categorical emotion simultaneously. | | | | The paper provides a comprehensive review of the state-of-the-art in USRs, sentiment analysis, and emotion recognition. It covers a wide range of topics, including different approaches to USR extraction, different methods for applying USRs to sentiment analysis and emotion recognition, and the challenges and limitations of the current state-of-the-art. | |
| **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 first step is to extract USRs from the speech signal. This can be done using a variety of approaches, such as deep learning-based approaches and traditional signal processing-based approaches. | USRs are designed to be universal, meaning that they can be used for a variety of speech processing tasks, including sentiment analysis and emotion recognition. This makes them a versatile tool for developing speech-based applications. | There is a lack of large-scale datasets that are labeled for both speech and sentiment or emotion. This can make it difficult to train and evaluate sentiment analysis and emotion recognition systems that use USRs. | | **2** | Once the USRs have been extracted, they can be applied to sentiment analysis or emotion recognition using a variety of methods, such as machine learning methods and deep learning methods. | USRs are typically more robust to noise and other variations in the speech signal than traditional speech features. This makes them more suitable for real-world applications. | USRs may not be as effective for sentiment analysis and emotion recognition in certain domains, such as medical speech or legal speech. This is because the speech characteristics in these domains may be different from those in general-purpose speech datasets. | |  |  |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | For each sentiment analysis and emotion recognition task, the accuracy scores can be considered dependent variables. | independent variable could be the number of classes in sentiment analysis tasks | A moderating variable could be the size of the datasets used for training the models. | A mediating variable could be the acoustic features extracted by the model, which serve as an intermediary step in the process of sentiment analysis and emotion recognition from speech. | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In the research paper, sentiment analysis and emotion recognition tasks are evaluated (dependent variables) using different UniSpeech-SAT models (independent variables). The dataset size may moderate performance, while acoustic features extracted by the models (mediating variables) contribute to understanding the complex relationship between model architecture and task outcomes. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | The input to the solution is a speech signal. The speech signal can be in any language and can be noisy. | The output of the solution is the sentiment or emotion of the speech signal. The sentiment can be positive, negative, or neutral. The emotion can be happiness, sadness, anger, fear, or surprise. | | | | The key feature of the solution is the use of universal speech representations (USRs). USRs are a type of speech representation that is designed to be universal, meaning that they can be used for a variety of speech processing tasks, including sentiment analysis and emotion recognition. USRs are typically more robust to noise and other variations in the speech signal than traditional speech features, and they have been shown to achieve state-of-the-art results on a variety of sentiment analysis and emotion recognition benchmarks. | | | | The solution has the potential to improve significantly the performance of speech-based applications, such as customer service chatbots, virtual assistants, and medical diagnostic systems. By using USRs, these applications can better understand the sentiment and emotion of the user, which can lead to more accurate and effective responses. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The solution can be used to develop customer service chatbots that can better understand the sentiment and emotion of the customer, which can lead to more accurate and effective responses. This can improve customer satisfaction and reduce the cost of customer service. | | | | | The solution can be used to develop virtual assistants that can better understand the user's intent and emotion, which can lead to more personalized and engaging experiences. This can increase user engagement and adoption of virtual assistants. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The paper is well-written and informative. The paper provide a good overview of the state-of-the-art in USRs, sentiment analysis, and emotion recognition. They also discuss the potential of USRs to improve significantly the performance of sentiment analysis and emotion recognition systems. However, the paper does not provide a detailed evaluation of the proposed solution, nor does it discuss the potential negative impacts of the solution in detail. | | | | The research paper on sentiment analysis and emotion recognition from speech using universal speech representations (USRs) does not provide a detailed evaluation of the proposed solution. However, the authors mention that they are planning to evaluate the solution on a variety of datasets in future work. | | | Abstract   1. Introduction 2. Related Work 3. Methods 4. Experiments 5. Results and Discussion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

**--End of Paper 4—**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Version 2.0 Week 2** | | | | | | | |
| **5** |
| **Reference in APA format** | |  | | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | | |
| https://www.ijert.org/stuttered-speech-recognition-using-convolutional-neural-networks | | Phani Bhushan S | | | | Stuttered Speech Recognition (SSR), Convolution Neural Network (CNN), Mel Frequency Co-efficient (MFCC). | | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The goal of the research paper is to develop a system that can recognize stuttered speech using convolutional neural networks (CNNs). CNNs are a type of machine learning algorithm that is well-suited for tasks such as image recognition and speech recognition.** | | | | **The paper use a variety of features to represent the stuttered speech signals, including Mel-frequency cepstral coefficients (MFCCs), pitch, and energy.** | | |
| Stuttered Speech Recognition using Convolutional Neural Networks | |  | | | |  | | |
| **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 stuttered speech data is prepared by performing data augmentation, noise injection, and masking. This helps to improve the robustness of the system to noise and other variations in the speech signal. | It achieves high accuracy in distinguishing between stuttered and non-stuttered speech. | It requires a large amount of data to train the CNN. | | **2** | A variety of features are extracted from the stuttered speech signals, including MFCCs, pitch, and energy. These features are used to represent the stuttered speech signals in a way that can be easily processed by the CNN. | It is robust to noise and other variations in the speech signal. | It is sensitive to the choice of hyperparameters, such as the learning rate and the batch size. | | **3** | The CNN is trained using the Adam optimizer and the cross-entropy loss function to distinguish between stuttered and non-stuttered speech. | It is computationally efficient and can be implemented in real time. | It is prone to overfitting, especially when trained on a small amount of data. | | **4** | The trained CNN is used to recognize stuttered speech in new data. |  |  | | **5** |  |  |  | | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Measurement of accuracy in stuttered speech recognition serves as the dependent variable, reflecting the effectiveness of the proposed system. | The independent variable is the implementation of the system using a combination of Weighted Mel Frequency Cepstral Coefficient feature extraction and Convolutional Neural Networks. | moderating variables such as the characteristics of the speech dataset, variations in speech patterns, or other contextual factors. | Identify any mediating variables that influence the relationship between the independent variable (SSR system) and the dependent variable (accuracy in stuttered speech recognition). | | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In this study, the independent variable is the implementation of a Stuttered Speech Recognition (SSR) system using CNN and Weighted MFCC. The dependent variable is the accuracy of stuttered speech recognition. Contextual factors, such as variations in speech datasets, act as moderating variables influencing system effectiveness. The underlying processes in the SSR system constitute mediating variables explaining the mechanism through which the system impacts accuracy. | | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** | |
| |  |  | | --- | --- | | **Input** | **Output** | | The input to the proposed stuttered speech recognition system is a stuttered speech signal. | The output of the system is a binary classification label, indicating whether the input speech signal is stuttered or non-stuttered. | | | | It uses a convolutional neural network (CNN) to extract features from the stuttered speech signal. CNNs are well-suited for this task because they are able to learn spatial and temporal patterns in the data. | | | | The system is robust to noise and other variations in the speech signal. This makes it suitable for use in real-world applications. | |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | | |
| The system can be used to develop new speech recognition systems that are specifically designed for stuttering speakers. These systems could be used to improve the performance of speech recognition systems in noisy environments, to make it easier for stuttering speakers to control smart devices, and to develop new communication tools for people who stutter. | | | | | **The system is prone to overfitting, especially when trained on a small amount of data. This means that the system may learn the characteristics of the training data too well and be unable to generalize to new data.** | | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** | |
| The research paper proposes a novel approach to stuttered speech recognition using convolutional neural networks (CNNs). The authors show that their proposed approach achieves high accuracy in distinguishing between stuttered and non-stuttered speech, even in noisy environments. | | | | Convolutional Neural Networks | | | 1. INTRODUCTION 2. LITERATURE SURVEY 3. PROPOSED METHODOLOGY 4. CONCLUSION | |
| **Diagram/Flowchart** | | | | | | | | |
|  | | | | | | | | |

**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** |
| **Bagus Tris Atmaja, Akira Sasou. 2022** | To create a deep learning-based model, particularly based on a time-delay neural network (TDNN), capable of accurately identifying various types of stuttering disfluencies solely from acoustic signals | There are two components, feature extraction and Evaluation | StutterNet utilizes TDNN to process MFCC features, capturing temporal aspects for accurate stuttering detection. | MFCC features extracted from speech signals, capturing stutter-specific characteristics for robust identification in StutterNet. | $1000 and $10000, for basic | TDNN and MFCC is fast |  | Proposed system provides good performance with accuracy: 74.07 | Acoustic signal reliance.  Few trainable parameters.  Promising stuttering detection results. | Limited research in stuttering.  Relatively small dataset size.  Context window optimization challenges. |  | The results indicate promising stuttering detection with StutterNet, outperforming a competitive method in various disfluency types. |
| **Shakeel A. Sheikh, Md Sahidullah . 2021** | The specific aim is to achieve accurate stuttering detection using a time-delay neural network (TDNN) architecture with Mel-frequency cepstral coefficients (MFCCs) | There are two components TDNN architecture, MFCCs | StutterNet uses TDNN to capture temporal context, enhancing acoustic features for accurate stuttering detection. | Mel-frequency cepstral coefficients (MFCCs), capturing stuttering nuances, crucial for StutterNet's acoustic analysis. | $1000 and $10000, for basic | TDNN and MFCC is fast |  | Proposed system provides good performance with accuracy: 80.01 | Accurate stuttering detection.  Efficient acoustic feature utilization.  Outperforms existing methods. | Limited evaluation on complex scenarios.  Reliance on a single dataset.  Optimization requires careful consideration. |  | The research paper demonstrates that StutterNet outperforms existing methods, achieving a 4.69% gain in accuracy and a 0.03 increase in Matthew's correlation coefficient (MCC). |
| **Colin Lea, Zifang Huang. 2023** | To address the challenges faced by people who stutter (PWS) in using speech recognition technology. | The components are: ASR models, dysfluency refinement, endpoint tuning | ASR tuning, endpoint adjustments, and dysfluency refinement; mechanisms improve speech recognition for people with stuttering. | Endpoint tuning, ASR model adjustments, and dysfluency refinement enhance speech recognition for individuals with stuttering. | $1000 and $10000, for basic | ASR can process data fast |  | Proposed system provides good performance with accuracy: 79.01 | Improved stuttered speech recognition.  Minimal data requirements.  Significant error rate reduction. | Some potential trade-offs.  User preferences may vary.  Limited context variability analysis. |  | what is the result of the research paper |
| **Phani Bhushan S, Vani H Y. 2021** | Stuttered Speech Recognition (SSR) system using Convolutional Neural Networks (CNN) and Weighted Mel Frequency Cepstral Coefficients. | The components are CNN , and Mel Frequency Cepstral Coefficients | Extracts features from stuttered speech using Weighted MFCC, classifies with CNN, enhancing speech recognition accuracy. | Weighted MFCC, CNN architecture, convolutional and pooling layers, achieving 92% accuracy in stuttered speech recognition. | $1000 and $10000, for basic | CNN can process data fast |  | Proposed system provides good performance with accuracy: 90 | Improved stuttered speech recognition.  Utilizes CNN for efficiency.  Achieves high accuracy rate. | Some potential trade-offs.  User preferences may vary.  Limited context variability analysis. |  | The research paper achieves a 92% accuracy rate in recognizing stuttered speech using the proposed Stuttered Speech Recognition (SSR) system with Convolutional Neural Networks (CNN) and Weighted Mel Frequency Cepstral Coefficients (MFCC). |
| **Suwon Shon , Pablo Brusco. 2021** | The primary goal of the research paper is to investigate and demonstrate the effectiveness of leveraging pre-trained language models, particularly BERT, in the context of speech sentiment analysis. | ASR encoder, pre-trained language models (BERT), sentiment classifier, pseudo label-based semi-supervised training. | Embedding pre-trained language models in sentiment classifiers, enabling robust sentiment analysis in spoken language. | Two-step pipeline, end-to-end system, BERT-based models, pseudo label-based semi-supervised training, sentiment analysis metrics. | $1000 and $10000, for basic | ASR can process data fast |  | Proposed system provides good performance with accuracy: 81.02 | Enhanced sentiment analysis performance.  Reduced need for human annotation.  Efficient knowledge transfer from text. | Limited discussion on ASR updates.  Dependency on pre-trained LMs.  Lack of exploration on other text corpora. |  | Embedding pre-trained language models in sentiment classifiers, enabling robust sentiment analysis in spoken language.The paper highlights the effectiveness of leveraging pre-trained language models, particularly BERT, in improving sentiment analysis performance |

**© Dr Abeer Alsadoon 2016\_CSU Sydney Study Centre**