CS6611-Creative and Innovative Project

(Team-35)

Enhancing Audio Clarity through Advanced Disfluency Detection and Removal with CNN

Team Members:

Pooja M J (2021503532)

Manokar S (2021503028)

Bhumisvara K (2021503704)

Mentor:

Dr. B. Thanasekhar

INTRODUCTION:

Speech disfluencies refer to the interruptions or irregularities in the flow of spoken language that include silent pauses, prolongations, repetitions of words, hesitation markers like "um" and "uh. Efficient communication relies on smooth and uninterrupted speech. In addressing these interruptions, especially for fast-paced speakers, strategic inclusion of silent pauses becomes essential.

This project aims to tackle the issue of interruptions in spoken language using advanced technologies. The main goal is to detect and eliminate the interruptions, to incorporate the pauses whenever needed and ultimately enhancing the fluidity of spoken language for improved communication and processing tasks.

PROBLEM STATEMENT:

The frequent occurrence of disfluencies poses a significant challenge for effective communication processing. These interruptions disrupt the fluency, lead to the misinterpretation of the speaker’s intended message and degradation in overall audio quality. Additionally, accommodating rapid speech require the strategic inclusion of silent pauses in the audio content.

In such instances, there is a need for the detection and removal of these disfluencies in the recorded audio to enhance the overall speech quality and also analyzing the speaker’s delivery of speech for improving their communication skills.

OBJECTIVES:

1) Improve the overall quality and fluency of the spoken content through precise detection and efficient removal of disfluencies.
2) Inclusion of silent pauses to accommodate speakers with fast-paced delivery.
3) Ensure that the removal of such disfluencies does not affect the intended meaning or conveyance of the actual message.
4) Perform detailed analysis by comparing the audio content, emphasizing speaker’s speech rate to assess improvements effectively.
5) Implement a question pool generation system based on the audio content.

LITERATURE SURVEY:

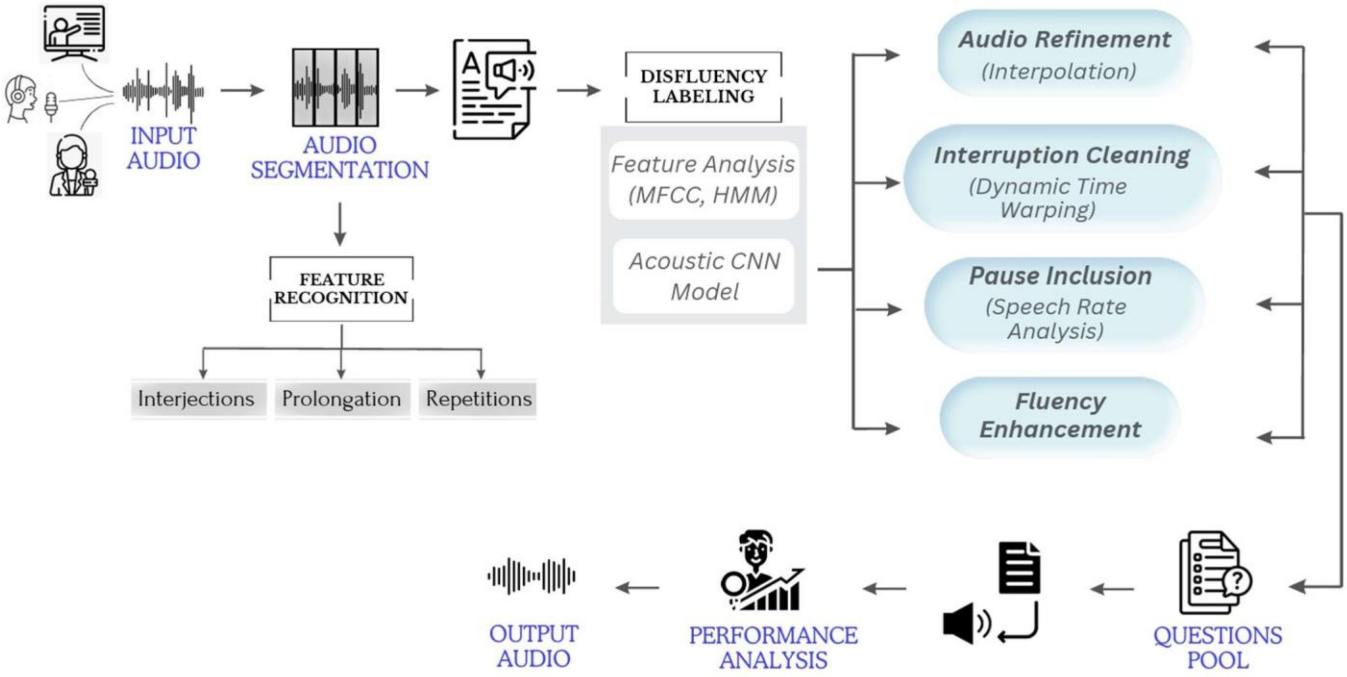
S. No	Name of the paper and published year	Methodology	Limitations
1)	FluentNet: End-to-End Detection of Stuttered Speech Disfluencies with Deep Learning Tedd Kourkounakis, Amirhossein Hajavi, Ali Etemad. IEEE Transactions, 2021	This paper involves the design of FluentNet, an automated stuttered speech detection. The network focuses on learning spectral frame-level representations and uses mechanism to better focus on the necessary features for stutter classification.	The study faces limitation in datasets and minimal impact from additional data. The paper focuses on binary detection (fluent / disfluent) but doesn't explore fine-grained classification of different disfluency types.
2)	Convolutional Neural Networks for Speech Recognition Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu IEEE Transactions, 2014	The methodology includes the use of hybrid DNN-HMM framework, description of basic CNN, proposal of limited-weight-sharing scheme, experiments on CNNs with FWS (Full Weight Sharing) and LWS (Limited Weight Sharing), and the use of a bigram language model in decoding.	The study lacks cross-cultural validation, raising uncertainty about the model. Further, the emphasis on classification over speech recognition limits insights into its broader applicability in Automatic Speech Recognition (ASR) tasks.
3)	End-to-End Neural Based Modification of Noisy Speech for Speech-in-Noise Intelligibility Improvement Muhammed P.V. Shifas, Yannis Stylianou IEEE Transactions, 2022	This includes the development of an end-to-end neural network approach for improving speech intelligibility in noisy conditions, training of both causal and non-causal FFTNet versions and utilization of a Teacher-Student approach	Existing approaches limitations with noisy speech highlight the need for robustness in real-world scenarios. Enhancements for real-time capability and efficiency are crucial, while potential integration with ASR systems could amplify recognition accuracy for automated processes.

4)	Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using Audio–Visual Information Serdar Yildirim, Shrikanth Narayanan IEEE Transactions, 2009	This paper includes a hidden event language model, prosodic cues, and innovative techniques like hidden-event part-of-speech modeling, multipass linear fold algorithm for sentence boundary detection, and various feature selection and classification methods for comprehensive disfluency detection.	The paper's limitations include the reliance on pitch breaks for candidate disfluent boundaries, potentially affected by the choice of the time interval, and the use of a simple approach for gestural information without explicit modeling of dynamic gestures.
5)	A CNN-Based Automated Stuttering Identification System Yash Prabhu, Naeem Seliya IEEE Conference, 2022	The methodology used in the study involved creating a CNN-based model, training and testing it using the Sep-28k dataset, using different training validation test splits, and measuring accuracy, recall, and precision using a confusion matrix.	Limited comparative analysis with other stuttering prediction systems, dataset specificity to English may hinder performance across diverse languages. Model enhancement includes integrate LSTM (Long-Short Term Memory) layers for sequential pattern recognition and refine dataset with precise label start and end points.
6)	Audio and ASR-based Filled Pause Detection Aggelina Chatziagapi, Dimitris Sgouropoulos, Constantinos Karouzos, Shrikanth Narayanan IEEE Conference, 2022	This paper introduces a framework for filled pause detection, utilizing both audio and textual data with separate experiments. It employs a CNN architecture for audio and a novel approach for text classification, addressing challenges posed by ASR systems with non-zero word error rates.	Challenges in handling disfluencies by ASR systems, difficulties in annotating disfluencies accurately, limited availability of annotated corpora, and the proposal of semi-automatic annotation as a solution.

7)	Improved Features and Models for Detecting Edit Disfluencies in Transcribing Spontaneous Mandarin Speech Che-Kuang Lin, Lin-Shan Lee IEEE Transactions, 2009	This involves developing features specific to Mandarin edit disfluencies and proposing an improved model combining decision trees and maximum entropy to detect IPs. The techniques were verified through experiments on a spontaneous Mandarin speech corpus.	Challenges in accurately classifying various edit disfluency types and balancing computational complexity and detection accuracy remain significant hurdles for real-time application feasibility.
8)	Edit disfluency detection and correction using a cleanup language model and an alignment model Jui-Feng Yeh, Chung-Hsien Wu IEEE Transactions, 2006	The paper employs acoustic features and Gaussian Mixture Model (GMM) for IP detection, incorporating hypothesis testing. It implements a two-stage disfluency correction module and evaluates performance on the Mandarin Conversational Dialogue Corpus.	Precise IP detection via acoustic features is hindered by the current approach's reliance on linguistic properties. Overcoming challenges related to subwords and enhancing restart disfluency performance is crucial.
9)	Recognizing Disfluencies in Conversational Speech Matthew Lease, Student Member, IEEE, Mark Johnson, and Eugene Charniak IEEE Transactions, 2006	The methodology involves utilizing a stochastic Tree Adjoining Grammar (TAG) noisy-channel model to generate candidate repair analyses, scoring fluency using a probabilistic syntactic language model, and selecting the most likely analysis with a maximum-entropy model.	The paper's disfluency detection, utilizing a stochastic Tree Adjoining Grammar model, is challenged by syntactic ambiguity and potential limitations in capturing nuanced contextual variations.

10)	Generation of Multiple-Choice Questions From Textbook Contents of School-Level Subjects Jianxing Yu, Qinliang Su, Xiaojun Quan IEEE Transactions, 2023	This paper involves the development of a system for automatic generation of MCQs using a pipeline with four primary modules: preprocessing, sentence selection, key selection, and distractor generation.	The system's domain-specific features may need adjustments for adaptation to non-educational contexts. Adaptation to more complex multiline facts and different subjects may require modifications to individual modules for optimal performance

ARCHITECTURE DIAGRAM:



CHALLENGES IDENTIFIED:

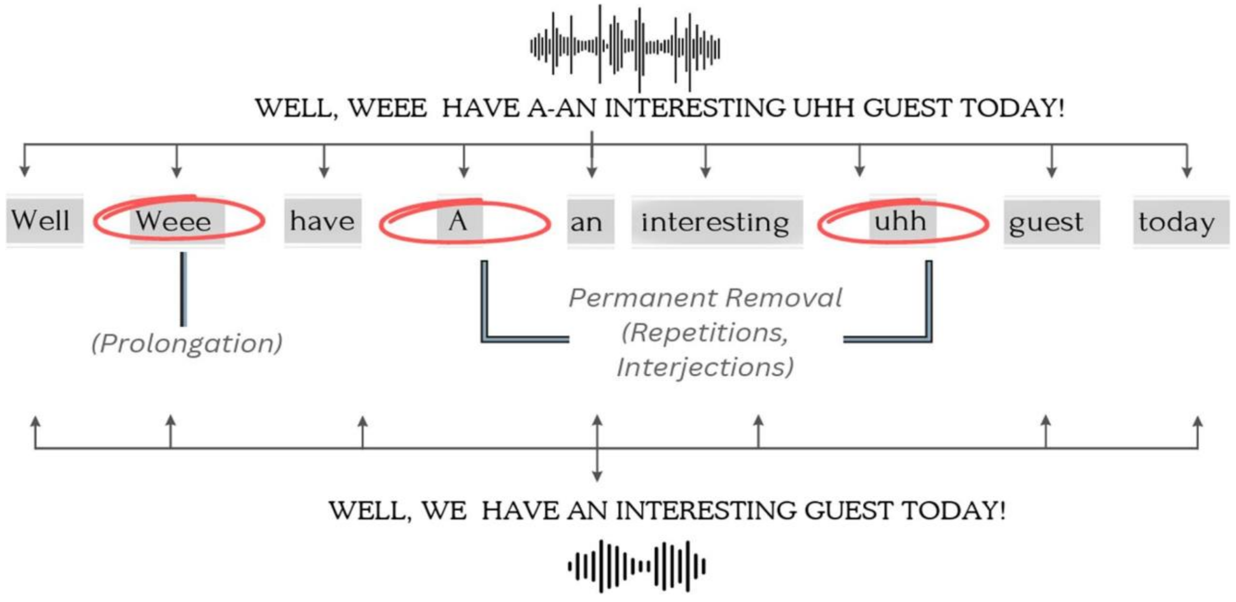
- **Accuracy Recognition**: Achieving high accuracy in recognizing different types of speech disfluencies is crucial.
- **Impact on naturalness:** Balancing the removal of disfluencies and preventing unintended alterations may affect the overall communication experience.

- **Data Scarcity and Annotation**: Labelling disfluencies is time-consuming and subjective.
- **Fast-paced Speech Handling**: Addressing the need for pause inclusion in the audio when a speaker's voice becomes too fast.

PROPOSED WORK:

- To utilize Convolutional Neural Networks (CNN) for analyzing the audio signals and addressing interruptions, particularly for speakers with fast-paced speech.
- Mel-frequency cepstral coefficients (MFCCs) or other features will be extracted from the preprocessed audio to capture speaker characteristics and speech patterns.
- Train the model and optimize the accuracy in detecting various disfluencies.
- For fast speakers, pauses will be inserted between speech segments based on the predicted disfluency locations and speaker's speaking rate.
- Remove the detected disfluencies using interpolation, noise reduction, or dynamic time warping and ensure that the removal process maintains the naturalness and authenticity of the speaker's voice.
- An interactive question pool will be generated based on the audio content and also performance of the speaker will be evaluated by analyzing the speech rate.

ILLUSTRATION:



MODULES:

1 – Data Preparation and Feature Analysis

2 – Training the Model

3 – Disfluency Detection

4 – Inclusion of Pauses

5 – Disfluency Elimination

6 – Questions Pool Generation

7 – Speaker’s Performance Analysis

Module 1:

Data Preparation:

- Collect a diverse set of labelled audio recordings, ensuring a balanced representation of fluent and disfluent speech.
- Organize the data into appropriate categories for training and testing purposes.

I) Audio Segmentation:

- The audio segmentation process involves dividing the raw audio data into smaller frames or segments.
- Each segment typically corresponds to a short duration of speech in the main dataset.
- The segmentation allows for the extraction of features from each frame, which are then used for further analysis and classification.
- The frames are processed to capture acoustic characteristics that can help in identifying filled pauses and other disfluencies in speech.

II) Frame-based Analysis:

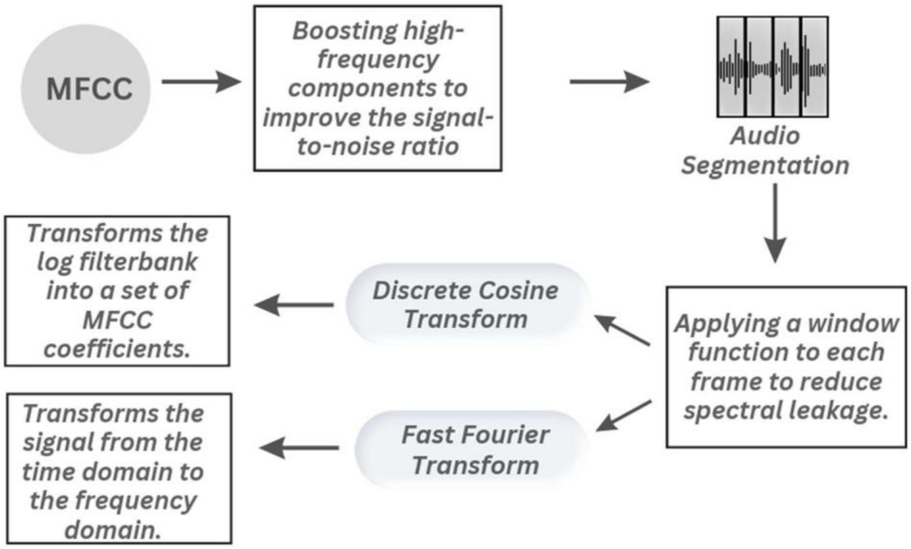
- By segmenting the audio data into frames and extracting relevant features, the system can analyze speech characteristics at a granular level.
- It enables the detection of filled pauses and other disfluencies within short segments of speech, contributing to more accurate classification results.

Feature Analysis:

- Acoustic features are extracted from each segmented audio frame to represent the characteristics of the speech signal.

I) Mel-Frequency Ceptral Coefficients (MFCC):

- MFCC is a widely used feature extraction technique and it involves Pre-emphasis, Framing, Windowing, FFT, Mel Filterbank, DCT.


MFCC
Boosting high-frequency components to improve the signal-to-noise ratio
Audio Segmentation
Applying a window function to each frame to reduce spectral leakage.
Discrete Cosine Transform
Transforms the log filterbank into a set of MFCC coefficients.
Fast Fourier Transform
Transforms the signal from the time domain to the frequency domain.

Algorithm : MFCC Feature Extraction

```
Input: Audio Signal
Output: MFCC Features
Function: extract(audio_signal, sample_rate, frame_length, frame_step)
Steps:
  1) Define a pre-emphasis coefficient.
  2) Apply a filter to the audio signal:
         pre_emphasized_signal = audio_signal[n] + alpha * audio_signal[n-1]
  3) Initialize empty list to store frames (frames = [])
  4) frame = pre_emphasized_signal[frame_size-1]
  5) Loop through each frame in the list:
       i.   windowed_frame = frame * window_function(len(frame))
       ii.  Compute the magnitude spectrum using DFT:
                magnitude_spectrum = abs(fft(windowed_frame))
       iii. Compute the DCT of the log mel energies:
                mfcc_coefficients = dct(log_mel_energies)
  6) End loop.
```

EVALUATION METRICS:

- Disfluency Detection Accuracy
- Fluency of the generated audio
- Overall speech quality improvement
- Speaker’s Performance Analysis

1) Pre-emphasis operation: **$y(t)=x(t)-\alpha\cdot x(t-1)$** where **$x(t)$** is the input signal, **$y(t)$** is the output signal, and **α** is the pre-emphasis coefficient.
2) The DFT formula is: **$X(k)$** = $\sum_{n=0}^{N-1} x(n). e^{-j2\pi kn/N}$ where **$X(k)$** is the DFT coefficient at frequency bin k, **$x(n)$** is the input signal, and **N** is the frame length.

REFERENCES:

1) T. Kourkounakis, A. Hajavi and A. Etemad, "FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning," in IEEE/ACM Transactions on Audio, Speech, andLanguage Processing, vol. 29, pp. 2986-2999, 2021, doi: 10.1109/TASLP.2021.3110146.
2) O. Abdel-Hamid, A. -r. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, Oct. 2014, doi: 10.1109/TASLP.2014.2339736.
3) M. P. V. Shifas, C. Zorilă and Y. Stylianou, "End-to-End Neural Based Modification of Noisy Speech for Speech-in-Noise Intelligibility Improvement," in *IEEE/ACM Transactions on Audio,Speech, and Language Processing*, vol. 30, pp. 162-173, 2022, doi: 10.1109/TASLP.2021.3126947. .
4) S. Yildirim and S. Narayanan, "Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using Audio–Visual Information," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 1, pp. 2-12, Jan. 2009, doi: 10.1109/TASL.2008.2006728.
5) Y. Prabhu and N. Seliya, "A CNN-Based Automated Stuttering Identification System," 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau,Bahamas, 2022, pp. 1601-1605, doi: 10.1109/ICMLA55696.2022.00247.
6) A. Chatziagapi et al., "Audio and ASR-based Filled Pause Detection," 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 2022, pp. 1-7, doi: 10.1109/ACII55700.2022.9953889.
7) C. -K. Lin and L. -S. Lee, "Improved Features and Models for Detecting Edit Disfluencies in Transcribing Spontaneous Mandarin Speech," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1263-1278, Sept. 2009, doi: 10.1109/TASL.2009.2014792..
8) Jui-Feng Yeh and Chung-Hsien Wu, "Edit disfluency detection and correction using a cleanup language model and an alignment model," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp. 1574-1583, Sept. 2006, doi: 10.1109/TASL.2006.878267.
9) M. Lease, M. Johnson and E. Charniak, "Recognizing disfluencies in conversational speech," inIEEE Transactions on Audio Processing, vol. 14, Sept. 2006, doi: 10.1109/TASL.2006.878269.
10) D. R. CH and S. K. Saha, "Generation of Multiple-Choice Questions From Textbook Contents ofSchool-Level Subjects," in *IEEE Transactions on Learning Technologies*, vol. 16, no. 1, pp. 40- 52, 1 Feb. 2023, doi: 10.1109/TLT.2022.3224232.