RESEARCH ARTICLE OPEN ACCESS

Rehearse: AI based feedback on communication skills

Prithvi Kumar, Ashwin Kurup, Anurag Saraswat, Karan Sharma,
Dr. Dashrath Mane
Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, Chembur, Mumbai, Maharashtra.

ABSTRACT
Being able to communicate effectively is perhaps the most important of all life skills. An AI-based platform is essential for honing your communication skills remotely. Many alternatives to this are available online, most notably BigInterview but almost all are hidden behind a paywall and/or are oriented towards interviewers. Hence these systems restrict themselves to only specific questions that the users must answer to, however, our system aims to simply analyze the speech of the user no matter what he/she may be speaking. In other words, the content of the speech does not matter. Our system aims to use all three aspects: Video, Audio and Text for analysis. Feedback is presented to the user in a web page along with a report that can be downloaded. Resources are also provided to the user to help with honing said communication skills, available on a separate page.

Keywords - Communication Skills, Personalized Feedback, Video processing, Text processing, Performance Parameters.

Date of Submission: 08-03-2023 Date of acceptance: 21-03-2023

I. INTRODUCTION

The ability to communicate information accurately, clearly and as intended, is a vital life skill and something that should not be overlooked. It's never too late to work on your communication skills and by doing so, you may well find that you improve your quality of life.

Our system involves an AI-based platform designed for honing your communication skills.Communication feedback is an essential aspect of self-reflection to further improve performance in job interviews as well.

II. LITERATURE REVIEW

We specifically looked into papers that focused on audiovisual detection and labeling to help us move forward in the project. Sushovan Chanda, Kedar Fitwe, Gauri Deshpande, Bjorn W. Schuller and Sachin Patel (2021)[1] looked at interviews conducted for 34 candidates, assessed their audio visual data and labeled them with varying levels of confidence. There exists higher confusion in detecting "high" and "low" confidence levels as compared to the "medium" confidence level. This paper showed the usage of different models and the process of labeling videos and audio samples on varying levels of confidence using various facial and verbal cues.

Tanvi Dinkar, Ioana Vasilescu, Catherine Pelachaud and Chloé Clavel (2020)[2] looked at the prediction of the perceived confidence of a listener from a speaker. They explored the importance of fillers in the prediction of confidence or perceived confidence of listeners, most of this work remains unexplored but gives a solid grounding on using fillers as a potential parameter for assessing communication skills.

Eye contact detection is an essential aspect of this project. In a paper with around 103 subjects, Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L. Ajodan, Melanie R. Silverman, Catherine Lord, Agata Rozga, and Rebecca M. Jones & James M. Rehg (2020)[3] developed a deep neural network model to automatically detect eye contact in egocentric video. They achieved a near equal accuracy to human experts.

This work[3] is the first of its kind to explore eye contact the way it has, and has proved as a solid grounding for building other neural networks to assess eye contact through Artificial Intelligence. Since our project looks at audio as well, collecting info from these kinds of papers was absolutely necessary. The project by Shruti Nair, Madhumita Mohan, Jemima Rajesh, and Priya Chandra (2020)[4] involved the building and scoring of an unbiased dataset of audio recordings based on the confidence of the speaker.

Since this paper[4] involved the comparison of three different models, out of which CNN proved to be the most accurate, also the usage and labeling of an unbiased dataset manually collected highlights the process very clearly for projects to undertake these steps and to create and label their own dataset.

For detection of filler words, a labeled dataset is absolutely necessary, for this purpose we found a paper by Ge Zhu, Juan-Pablo Caceres, and Justin Salamon (2022)[5] that provides solid info into the method for collecting such a dataset. They have collected over 35k filler words from podcasts and annotated them usefully.

We must also label audios to detect correct confidence levels, for this purpose Xiaoming Jiang and Marc D. Pell (2021)[6] provided a solid grounding. In this paper[6], the presence of a linguistic cue tended to increase ratings of confident voices but decrease ratings of voices in the less confident voice conditions. They explored the usage of positive, negative, and neutral linguistic cues, to figure out how these tended to change the perception of confidence among listeners.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J.Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu (2019)[7] demonstrated that the use of pre-trained Machine Translation models is more effective at converting weakly supervised learning data for Speech Translation. Since there is an unavailability of large amounts of data and their translated transcript pairs, this paper[7] showed that pre-trained MT models underperform when compared to ST models for Speech translation from unlabeled or largely unlabeled data.

Christine Dewi, Rung-Ching Chen, Xiaoyi Jiang, and Hui Yu (2022)[8] provided a novel method for blink detection that takes into account various factors that could affect the detection such as lighting conditions, facial emotions, and head position. According to the results from a typical data set, the suggested approach is more efficient than the state-of-the-art technique. Blink detection is an important technique in a variety of settings, including facial movement analysis and signal processing. This method provides a foundation for building methods for more accurate blink detection.

III. PROPOSED SYSTEM

The Proposed system aims to help users with their communication skills by providing them with a detailed analysis of their speech. The system takes video input from the user and extracts the required information from the provided video and performs the needed analysis on the extracted information to produce results for the user. Below are the steps in which the system is carrying out the whole process.

3.1 Input

Users must upload a recorded video of themselves speaking in English in order to receive feedback on their spoken communication. To get the best results from the various AI models utilized in the system, this video needs to be filmed in a quiet location with good lighting, where the speaker can be seen well and there is little to no background noise. This video will be used to extract data in appropriate formats to be used for further analysis such as audio and text format.

3.2 Processing

3.2.1 Audio processing

First the Audio is extracted from the provided video input and then further analysis is performed on that. We use speech recognition to detect the user’s pace, as well as the number of powerful words they have used. Audio can be used to identify if the person is speaking confidently; this can be done by analyzing the speaking tone from the audio and identifying stutters and random pauses.

3.2.2 Visual processing

Video input is analyzed to find out how often the person is blinking and eye contact is maintained by the person while speaking. This gauges the user’s confidence and comfort level in communicating their views.

3.2.3 Text processing

First the text is extracted from the video input provided by the user using Speech to Text engines that provide good accuracy. This text is used to derive various insights from it. We find the words per minute spoken by the user from the extracted text that gives the user an idea on how fast they are speaking and is it in a suggested range. We also identify and count the number of Power Words used by the user in a particular speech. Using Power Words in a speech is considered to be a sign of confidence. In future we can also check for grammatical mistakes and vocabulary used in speech through text analysis.

3.2.4 System output

After all calculations are completed, a REST API is used to send the results to the frontend, where users can view the results in a web UI. Text output that provides explanations of the results is supplied to the frontend along with the calculated results and graph data. The UI will allow users to view speech evaluations for themselves. These evaluations are displayed in the form of various graphs or in the most suitable format, and each result is accompanied by a thorough explanation of what can be deduced from it.

IV. METHODOLOGY AND IMPLEMENTATION

Spoken Communication as a soft skill is highly valued in today's job market. Significant number of people struggle while communicating ideas especially when speaking in English
This project provides a solution for this very problem by aiming at evaluating spoken communication of users via a deep learning based web application. Figure 1 shows the basic working of the system, as elaborated below.

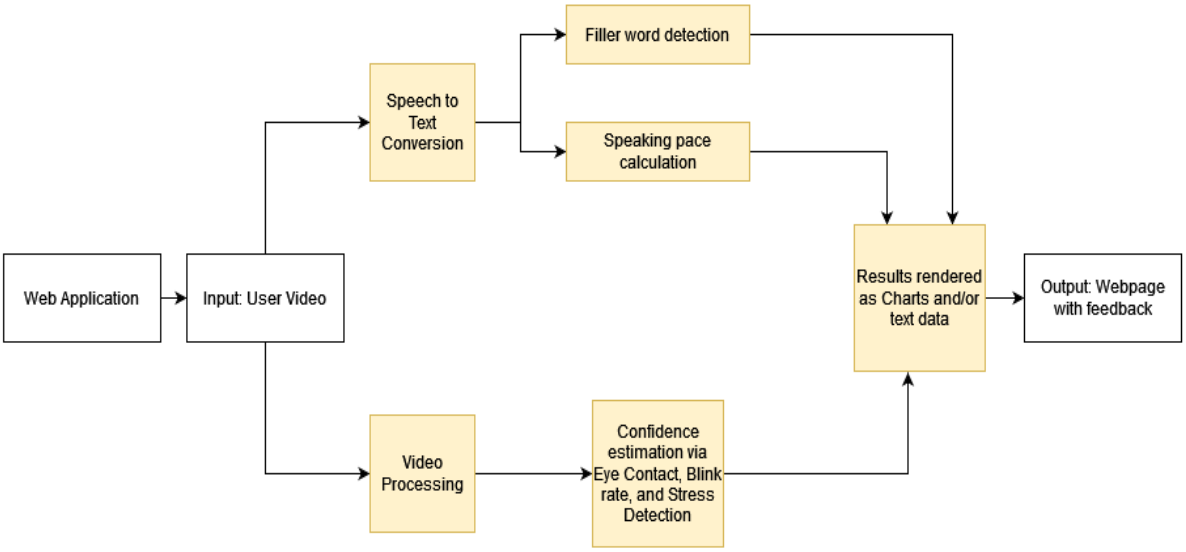

Web Application
Input: User Video
Speech to Text Conversion
Filler word detection
Speaking pace calculation
Video Processing
Confidence estimation via Eye Contact, Blink rate, and Stress Detection
Results rendered as Charts and/or text data
Output: Webpage with feedback

Fig 1. *Modular diagram of the proposed system*

The video input will be extracted into two domains for analysis
- Audio
- Text via speech processing

The methodology behind the server-side working of the system can be broken down into the following parts:

i. **Eye contact**: To calculate the amount of eye contact maintained by the user, a Dlib model is used to isolate 68 landmarks on the face. From these, 12 points are used to detect the eyes and pupils. OpenCV iterates over each frame of the video and passes it to the model, which returns whether the user is looking at the camera, or looking around. Variables are used to record the eye contact status at each frame, and the data is passed to the frontend for graph generation.

ii. **Blinks per minute**: The above model also detects blinking, however one blink can go across multiple frames. Consecutive frames of blinks are grouped together and the number of blinks is counted overall. This is then passed to the frontend as well.

iii. **Pace**: The audio is extracted from the video and passed to Google's speech recognition API that accounts for world accents. The number of words is divided by the durations to give the pace in Words Per Minute.

iv. **Power Words**: The text obtained in the previous step is matched with a list of power words in order to determine which words the user has utilized, and thereby also giving the number of power words used. Both quantities are passed to the frontend.

v. **Overall Grading :** To enable the system to evaluate overall performance, we manually graded a set of sample videos into categories based on the speaker's performance. In addition, we use other parameters, such as eye contact, speaking pace, and stress level, to further inform the system's evaluation of the speaker's performance. By incorporating these parameters into the system's multi-layer perceptron model, the system is able to learn to grade overall performance based on the evaluated parameters. This enables the system to provide the user with an evaluation of their overall performance and confidence level, based on an informed analysis of their speaking abilities. By presenting users with a detailed evaluation of their speaking skills, the system can help them to identify areas for improvement and develop strategies to enhance their speaking abilities.

vi. **Pause detection :** For pause detection, we're using pydub, a python library that detects speech pauses and silences and labels them as pauses. We've chosen to show this to the user, only recording long pauses and frequent ones, as some are used in normal speech patterns devoid of stress. Pydub requires a lot of hyperparameter tuning as videos cannot always be of the highest quality.

The client-side working of the system can be broken down into the following parts:

i. **Graph generation**: The data passed from the server is used to generate graphs via plotly js for each quantity that we analyze.

ii. **Report generation**: For now the report generated is just a formatted version of the results page. Further down the line the goal will be to generate a more comprehensive report with a better layout that is personalized to each user.

iii. **Results page**: This page uses the graphs along with the results to display positive, neutral or encouraging messages based on the user's performance.

iv. **User Profile** : This feature we've added gives a personal layer to the whole project. Basically the user has to login with an email id and password and will have a profile to come back to anytime he/she wishes to check his/her history. The history includes the past grades, along with the personalized report given to them at each session they uploaded a video. The profile page implementation is shown in a figure in the results section.

v. **Personalized Report**: This feature is essentially an extension of the user's profile page. Earlier we simply generated a report based on the users performance that was deleted the moment the user clicked off the site, However, the reports now are saved and personalized for each session they sit in and can be accessed at the click of a button once the user has logged in, The evaluation is presented in a neat tabular form.

The implementation of the website at the current stage ends with the creation of three pages: Home. Upload, Results,Signup, Help Section and Profile page.

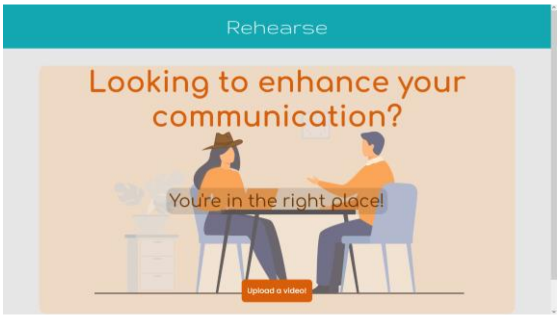
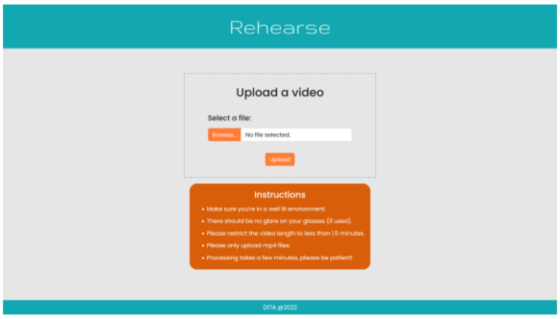

Rehearse
Looking to enhance your communication?
You're in the right place!
Upload a video!

Fig 2. *Homepage*

Figure 2 illustrates the homepage, which acts as a point of information, showing the user the purpose of the application and how they can use it.



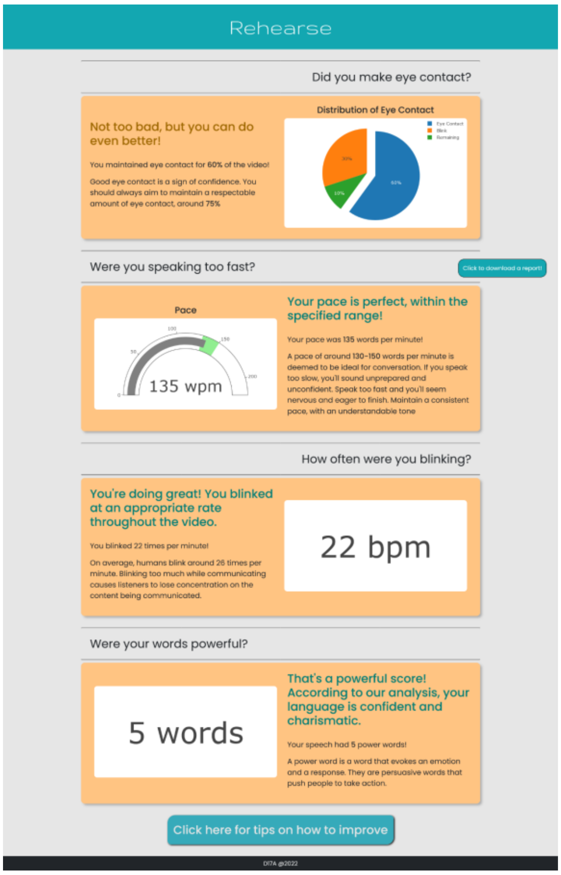
Instructions

- Make sure you're in a well lit environment.
- There should be no glare on your glasses (if used).
- Please restrict the video length to less than 1.5 minutes.
- Please only upload mp4 files.
- Processing takes a few minutes, please be patient!

Fig 3. *Upload page*

Figure 3 illustrates the upload page. There are a few instructions listed which elaborate the limitations of the system, and the file format required. Users can upload the video from their device. The uploaded video is processed, and then deleted from the server.

Figure 4 illustrates the results page. All the processing done on the server side is visualized here in the form of charts. The user can then understand their strengths and weaknesses. Additionally, they can follow the link at the bottom to find a few resources that can help them to improve those skills that are lacking.


Rehearse
Did you make eye contact?
Distribution of Eye Contact
Eye Contact
Blink
Remaining
30%
10%
60%
Were you speaking too fast?
Click to download a report!
Pace
0
50
100
150
200
135 wpm
How often were you blinking?
22 bpm
Were your words powerful?
5 words
Click here for tips on how to improve
DI7A @2022

Not too bad, but you can do even better!

You maintained eye contact for 60% of the video!

Good eye contact is a sign of confidence. You should always aim to maintain a respectable amount of eye contact, around 75%

Your pace is perfect, within the specified range!

Your pace was 135 words per minute!

A pace of around 130-150 words per minute is deemed to be ideal for conversation. If you speak too slow, you'll sound unprepared and unconfident. Speak too fast and you'll seem nervous and eager to finish. Maintain a consistent pace, with an understandable tone

You're doing great! You blinked at an appropriate rate throughout the video.

You blinked 22 times per minute!

On average, humans blink around 26 times per minute. Blinking too much while communicating causes listeners to lose concentration on the content being communicated.

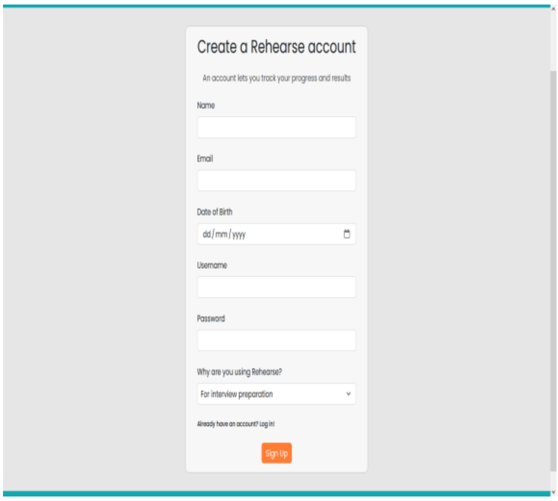
That's a powerful score! According to our analysis, your language is confident and charismatic.

Your speech had 5 power words!

A power word is a word that evokes an emotion and a response. They are persuasive words that push people to take action.

Fig 4. *Results page*

Figure 5 illustrates the signup page. It includes the email address, password,date of birth, name and the reason for using our platform. The last question is purely for research purposes to understand where our user base is coming from and what their needs are. Additionally, the entire thing adds a personal layer to the project.



Create a Rehearse account

An account lets you track your progress and results

Name

Email

Date of Birth
dd / mm / yyyy

Username

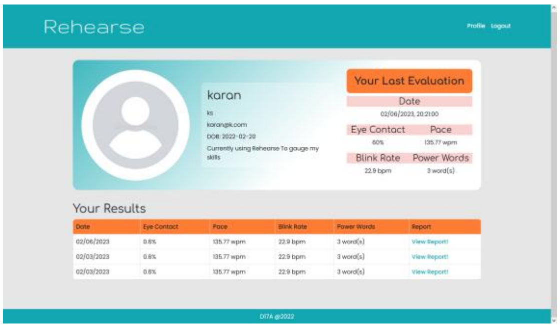
Password

Why are you using Rehearse?
For interview preparation

Already have an account? Log in!

Fig 5. *The signup page.*

Fig 6 shows the Profile section of the user. Once the user has logged in, the profile section will help the user keep track of their evaluation history and also give personalized reports for each session they sat on in the form of links and presented in a neat tabular form.



karan
ks
karan@k.com
DOB: 2022-02-20
Currently using Rehearse To gauge my skills

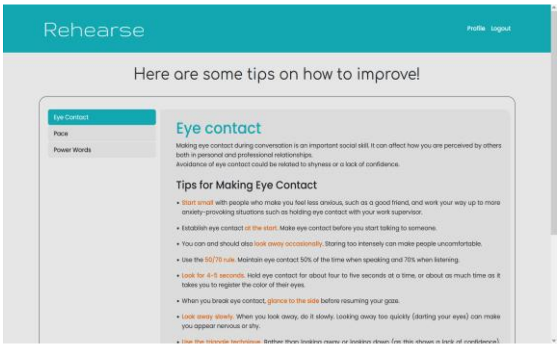
Your Last Evaluation
Date
02/06/2023, 20:21:00
Eye Contact 60%
Pace 135.77 wpm
Blink Rate 22.9 bpm
Power Words 3 word(s)

Your Results

Date	Eye Contact	Pace	Blink Rate	Power Words	Report
02/06/2023	0.6%	135.77 wpm	22.9 bpm	3 word(s)	View Report!
02/03/2023	0.6%	135.77 wpm	22.9 bpm	3 word(s)	View Report!
02/03/2023	0.6%	135.77 wpm	22.9 bpm	3 word(s)	View Report!

Fig 6. *The profile section.*

Figure 7 shows the helper section of our websites for all users. It basically has some tips to improve communication skills and cites some sources to help the user on each aspect or parameter that they feel the need to improve upon.



Here are some tips on how to improve!

Eye contact

Making eye contact during conversation is an important social skill. It can affect how you are perceived by others both in personal and professional relationships.
Avoidance of eye contact could be related to shyness or a lack of confidence.

Tips for Making Eye Contact

- Start small with people who make you feel less anxious, such as a good friend, and work your way up to more anxiety-provoking situations such as holding eye contact with your work supervisor.
- Establish eye contact at the start. Make eye contact before you start talking to someone.
- You can and should also look away occasionally. Staring too intensely can make people uncomfortable.
- Use the 50/70 rule. Maintain eye contact 50% of the time when speaking and 70% when listening.
- Look for 4-5 seconds. Hold eye contact for about four to five seconds at a time, or about as much time as it takes you to register the color of their eyes.
- When you break eye contact, glance to the side before resuming your gaze.
- Look away slowly. When you look away, do it slowly. Looking away too quickly (darting your eyes) can make you appear nervous or shy.
- Use the triangle technique. Rather than looking away or looking down (as this shows a lack of confidence),

Fig 7. *The helper section.*

V. RESULTS

The Video Processing model performs well in detecting eyes in frames with proper lighting but fails at times in insufficient lighting or in the case of users wearing spectacles, if there is any glare on the spectacles generating Lost Frames. These Lost Frames are excluded from all calculations, however a higher number of lost frames can result in inaccurate results and therefore it is necessary for the uploaded video to have proper lighting and no glare throughout.

The model is able to detect whether a frame has the user blinking or not. However, since a single blink can extend over multiple frames with different durations, the number of blink frames isn't sufficient enough to calculate the blink rate. Therefore, we count the number of groups of successive blink frames to find the number of blinks. This is divided by the duration of the video, adjusted for lost frames, to give the blink rate.

The other evaluation parameters have already been discussed above at length. The audio processing model also includes pause detection which helps users understand good and bad use of pauses and it is displayed on the user results section. It needed quite a bit of tuning to understand pause length and decibel value to help fine tune it to match our users input videos as they are not always of the highest quality.

Currently, we have implemented a personalized report to help the user keep a track of each session they sit in. This is presented in a neat tabular form in the users profile section . Existing features can be tuned and improved upon as well to give more accurate results to the users.

VI. CONCLUSION

The end product is a web application which outlines a convenient way for users to prepare and work on themselves.Our goal is bridge the gap between the existing communication and interview preparation methods; allowing for a more flexible and robust way to grow.

The development of this system shows how AI can be used to provide feedback to anyone based on their communication skills.Additionally, it highlights the importance of considering a wide range of parameters when it comes to communication assessment.

The application of various parameters like Video, Audio and Text on the user proves to be a useful and reliable tool for the hiring industry to depend upon.

Also remote communication assessment is a widely unexplored field in terms of research done, this project helps provide that necessary momentum to understand the potentiality and necessity of more research to be done here.

REFERENCES

[1]. Sushovan Chanda, Kedar Fitwe, Gauri Deshpande, Bjorn W. Schuller, Sachin Patel, A Deep AudioVisual Approach to Human Confidence Classification , Front. Computer. Sci., 2021

[2]. Tanvi Dinkar, Ioana Vasilescu, Catherine Pelachaud, Chloé Clavel,How confident are you? Exploring the role of fillers in the automatic prediction of a speaker’s confidence, ICASSP ‘20: International Conference on Acoustics, Speech and Signal Processing, 2020.

[3]. Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L. Ajodan, Melanie R. Silverman, Catherine Lord, Agata Rozga, Rebecca M. Jones & James M. Rehg, Detection of eye contact with deep neural networks is as accurate as human experts,Nature Communications, 2020.

[4]. Shruti Nair, Madhumita Mohan, Jemima Rajesh, Priya Chandra, Finding the Best Learning Model for Assessing Confidence in Speech, 2020, MLMI '20: 2020 The 3rd International Conference on Machine Learning and Machine Intelligence

[5]. Ge Zhu, Juan-Pablo Caceres, Justin Salamon, Filler Word Detection and Classification: A Dataset and Benchmark, University of Rochester, Adobe Research, Interspeech, 2022.

[6]. Xiaoming Jiang , Marc D. Pell ,Encoding and decoding Confidence information in speech, School of Communication Sciences and Disorders and Center for Research on Brain, Language and Music, McGill University, Canada, 2021.

[7]. Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J.Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, Yonghui Wu, Leveraging Weakly Supervised Data to improve End-to-End Speech-to-Text Translation, ICASSP ‘19,International Conference on Acoustics , Speech and Signal Processing, 2019.

[8]. Christine Dewi, Rung-Ching Chen, Xiaoyi Jiang, Hui Yu, Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks, PeerJ Comput Sci. ,2022.