Two-Way Sign Language Translator for Visually Impaired People

Ms. KAVITHA R[1], Ms. Varsha S[2], Mr. Bharaneedharan C[3]

[1]Assistant Professor Level – I, Computer science and engineering & Bannari Amman Institute ofTechnology
[2] UG Scholar, Computer Science and Engineering & Bannari Amman Institute of Technology
[3] UG Scholar, Electronics and Instrumentation Engineering & Bannari Amman Institute of Technology

---

Abstract – Hearing disabilities affect people all over the world. Nearly 360 million people have a disabling hearing loss. In the recent years, there has been a rapid increase in the number of deaf and dumb victims due to birth defects, accidents and oral diseases. Since deaf and dumb people cannot communicate with a hearing person, they have to depend on some sort of visual communication. Sign Language is not understood by the majority of hearing people. The objective of our project is to form a bridge between the hearing-impaired community and the hearing community; to further commence a two-way communication. We propose a two-way sign language translator that converts Voice to sign language and vice versa in real time. We process the video frame by frame using the OpenCV Library in Python. Furthermore, we use background subtraction method called Gaussian Mixture-based Background/Foreground Segmentation Algorithm on each frame to subtract the background. Contours of this processed image are then passed into a Deep Neural Net (DNN) to classify the frame to the corresponding written language equivalent words. For conversion from sign language to voice, we use basic Natural Language Processing and gtts to accurately conserve the grammar in the sign language. The enrolment rate and literacy among the hearing-impaired children is far below the average for the population at large. This technology will then help the normal schools better integrate the hearing-impaired community thus making education more accessible and cheaper for them. Speech Emotion Recognition (SER) is a vital aspect of human-computer interaction and emotional intelligence applications. This SER that employs Mel-frequency cepstral coefficients (MFCC) as the basis for feature extraction, followed by a Convolutional Neural Network (CNN) with a customized architecture for emotion classification. The methodology 2 involves transforming audio samples into MFCC images, enabling the application of CNNs typically used in computer vision tasks. This fusion of traditional audio analysis with deep learning techniques harnesses the power of MFCC's spectral representation and CNN's spatial pattern recognition. By modifying the CNN architecture, the model can better discern emotional cues in the MFCC images, enhancing SER performance. This strategy offers SER a viable path forward, with the ability to identify complex emotional expressions. In order to increase the accuracy of emotion recognition, it tackles the requirement for feature-rich representations and sophisticated modelling techniques.

KEYWORDS: *OpenCV, DNN, SER, CNN, MFCC*

1. INTRODUCTION

1.1 Two-way Sign Language: According to the World Health Organization, 72 million people worldwide are deaf and over 466 million have a debilitating hearing loss. Before the invention of sign language, the deaf-mute community faced significant communication barriers that hindered their advancement. For those who are deaf, sign language serves as both their primary means of communication and their fundamental form. Sign language is characterized as a language that combines body posture, facial expressions, and hand movements to convey meaning. Because of its significance, the International Day of Sign Languages is observed on September 23 by the UN, which promotes sign language as a human right and on par with spoken languages. Despite the fact that sign language is the major means of communication for the deaf-mute people, it is still extremely important. Since very few people outside of the deaf-mute population are familiar with sign languages, it presents a significant obstacle to communication between the speaking and deaf-mute communities and, consequently, to the advancement of both societies as a whole. Voice to sign language translators and sign language to voice translators are the two categories into which the current translators of sign

language fall. Typically, sign-to-voice methods rely on either image or sensor data, and occasionally they even combine the two. They have already demonstrated a concept that uses gloves with sensors integrated in them to identify the ASL signs that a deaf person makes. The application of deep neural networks to the analysis of spatiotemporal data for sign language recognition has gained popularity in recent years. For example, classify movements using the DCNN architecture. Despite its great accuracy, the DCNN architecture needs RGB-D (depth-dimensional) inputs, which come from devices that have depth sensors specifically designed for that purpose.

1.2 Speech Emotion Recognition (SER): Speech, being the principal means of human communication, not only transmits data and words but also a complex web of feelings. Understanding these emotional cues in speech and being able to identify them is a basic component of human connection. Our emotional moods have a big impact on how we convey and understand messages in both casual and customer service settings. The multidisciplinary field of speech emotion recognition (SER) lies at the nexus of signal processing, psychology, and computer science. Its goal is to create automated systems that can recognize, comprehend, and categorize the emotional content that is expressed in spoken language. Applications for these systems are numerous and include sentiment analysis, virtual assistants, mental health evaluation, and human-computer interaction. It is impossible to overestimate the significance of SER in the modern digital world since it is essential to raising the efficacy and efficiency of a wide range of apps and services. The human voice is a veritable gold mine of emotional data. Prosody—a term used to describe variations in pitch, tone, rhythm, and intensity—is essential to expressing emotions through speech. The emotional context is further enhanced by the speaker's voice quality, word choice, and delivery. While humans are naturally good at identifying these signs, it is a difficult and constantly changing task to teach robots the same skills. This introduction provides a thorough review of the area of SER, emphasizing its importance, uses, difficulties, and most recent developments. It also explores the methods used in SER research as well as the psychology that underlies speech emotion.

2. LITERATURE REVIEW

1.) Title: Speech Emotion Recognition Using Deep Learning Techniques.

Authors: Harshitha, Janumpally Sushma, Namsamgari Mukesh.

Publication: IEEE Access Recognizing emotions from voice signals is a crucial but difficult aspect of HCI (human-computer interaction). Numerous methods, including well - known speech analysis and classification approaches, have been used in the literature on speech emotion recognition (SER) to extract emotions from signals. Recently, deep learning approaches have been put out as a substitute for conventional methods in machine learning. This paper examines some current literature that uses Deep Learning approaches for speech-based emotion recognition and provides an overview of these techniques. The review discusses the databases used, the emotions that were retrieved, the advancements made in speech emotion recognition, and its limits."

2.) Title: Speech Emotion Recognition using Machine Learning.

Authors: Kotikalapudi Vamsi Krishna,Navuluri Sainath,A. Mary Posonia.

Publication: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC) Finding the feelings that the speaker elicits while speaking is the paper's goal. These days, detecting emotions is a crucial task. Pitch ranges are higher and wider for speech in fear, wrath, and joy, but they are lower for speaking in other emotions. Speech recognition is helpful for enhancing human-machine communication. Here, we are identifying the emotions using several categorization techniques. The audio features MFCC, MEL, chroma, and Tonnetz were employed, together with Support Vector Machine and Multilayer Perception. These emotions—calm, neutral, surprise, pleased, sad, furious, afraid, and disgusted—have been programmed into these 13 models. We tested it using the input audio and obtained an accuracy of 86.5%."

3.) Title: Speech Emotion Recognition using Machine Learning

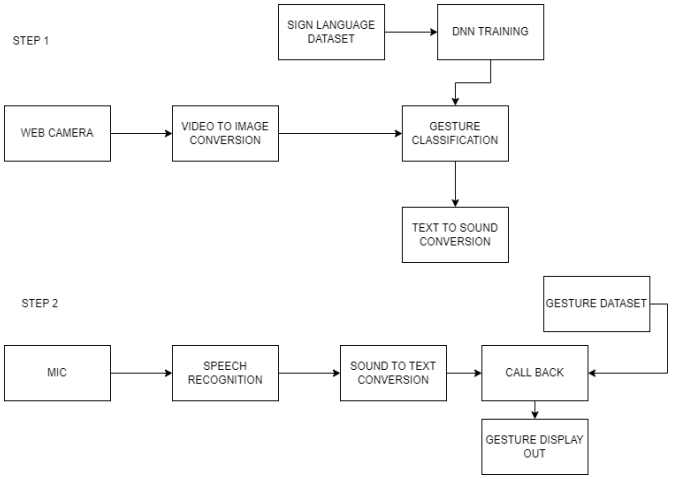
Authors: R. Anusha, P.Subhashini, Darelli Jyothi.

Publication: 2021 5th International Conference on Trends in

Electronics and Informatics (ICOEI) Speech emotion recognition is the process of accurately anticipating a person's emotion from their speech. Better human-computer interaction is produced by it. Speech Emotion Recognition (SER)u201d makes it possible to predict a person's emotion, even if emotions are subjective and audio annotation is difficult 1. Animals that are able to comprehend human emotion, such as dogs, elephants, horses, etc., use the same premise. 1. Many states, such as tone, pitch, expression, conduct, etc., can be used to forecast an individual's emotion. A select few of them are thought to be able to convey emotion through speaking. To train the classifiers for speech emotion recognition 2, a small number of examples are used. This research work considers the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song dataset). Here, the three key features such as MFCC (Mel Frequency Cepstral Coefficients), Mel Spectrogram and chroma are extracted."

3. PROPOSED WORK

SYSTEM DESIGN – TWO-WAY TRANSLATOR

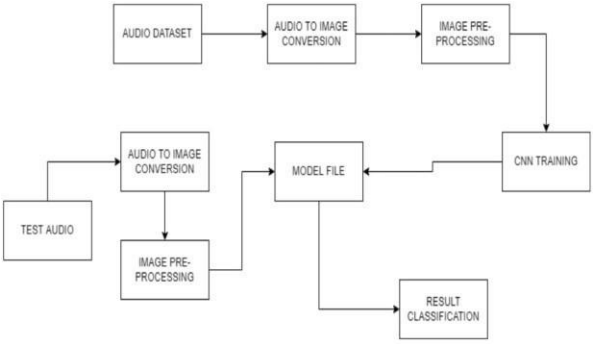


HAND GESTURE TO VOICE: A hand gesture recognition system is a useful piece of modern technology that is utilized for both disabled people and human-computer interaction. Various technologies and tools have been employed by the system to accomplish its objective. In-depth research and study are done on the tools and technologies as well as the earlier technologies that were employed. OpenCV: OpenCV is a collection of freely available computer vision libraries. Computer vision is the process of teaching machines intelligence and mimicking human vision. The opencv image processing library was developed by Intel, funded by Willow Garage, and is being kept up to date by idseez. It can be downloaded for Linux, Windows, and Mac. It works in C, C++, Python. It is open source and free as well as easy to use and install. OpenCV helps the deep learning frameworks like TensorFlow, Torch/PyTorch and Caffe. Vision in computers: Computer vision is the process of imparting intelligence to machines so they can perceive images or videos and gather data from them in a manner similar to that of humans. Writing computer programs that can interpret images or videos—that is, programs that can take images as input and output descriptive information about the visuals—is the aim of computer vision research. The image is first taken as an input, and after that, processing takes place to complete the process. Following that, the image is examined using the program's dataset, and the output—a feature extraction process that yields an image description—is displayed. The outcome may take any shape, including symbolic or numerical data. Convolutional neural network: A sort of deep neural network used in deep learning a convolutional neural network (CNN) works with a set of data to extract information about that data. Similar to how sounds, films, or photos can be used to extract data from CNN. CNN mostly consists of three items. Local receptive field comes first, followed by shared weight and biases, and activation and pooling in last order. To enable CNN to extract features from a particular input, the neural networks are first trained on a large amount of data. As soon as the input is received, pre-processing of the image is completed, followed by feature extraction from the collection of stored data, data classification, and output display. Only input for which the neural network has been trained and stored can be processed by CNN. Natural language processing, recommender systems, image classification, medical image analysis, and image and video recognition all employ them. Inception version three: One machine learning technique that uses a neural network that has already been trained is called transfer learning. This approach is primarily divided into two sections. First, a convolutional neural network is used for the feature extraction phase, and then a fully-connected network is used for the classification phase. For instance, the two components of the image recognition model known as Inception-v3 are the classification layer and the softmax layers.

VOICE TO HAND GESTURE: Using a Microphone for Speech Input and Text Translation Permit Ambient Noise Adjustment: Because ambient noise fluctuates, we need to give the software a little time to modify the recording's energy threshold so that it reflects the level of outside noise. Google Speech Recognition is used to translate spoken words into text. For this to function, you must have a live internet connection. Certain offline recognition systems, such PocketSphinx, do exist, but they require a complex installation process with multiple dependencies. Google Speech Recognition is among the most user-friendly. Speech to Text Translation: To begin, import the library and use the init() function to initialize it. This function has two possible inputs. init(driverName string, debug bool) drivername: [Name of available driver] sapi5 on Windows | nsss on MacOS

debug: to enable or disable debug output After initialization, we will make the program speak the text using say() function. This method may also take 2 arguments. say(text unicode, name string) text: Any text you wish to hear. name: To set a name for this speech. (optional) Finally, to run the speech we use runAndWait() All the say() Once the Speech is detected in mic it is then converted in to text and respective gestures are displayed according to conditions.

SYSTEM DESIGN – SER



CNN Architecture Changes A modified convolution architecture is created here in order to shorten the training time without sacrificing any significant features. The model consists of three convolution layers and two hidden layers. From the input, the image resolution is converted to 200 by 200 and sent to the first convolution layer. In each of the three convolution layers, the 3*3 kernel is activated with relu activation, which interprets the positive part of its argument after relu. The pooling process is then carried out; in our case, max-pooling is carried out following the three convolution layers, the two hidden layers with a dense of 128 is reduced, and softmax activation is carried out. Lastly, the created architecture is trained using the provided dataset and the CNN model file is created for classification.

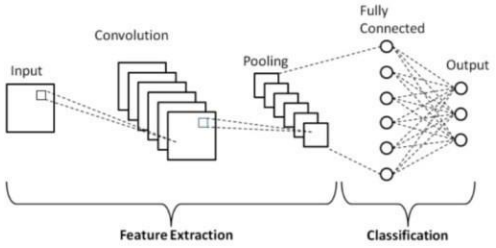


Fig.1 DenseNet Architecture

4. RESULTS:

The suggested solution uses software-based artificial intelligence approaches to lower hardware costs. Artificial intelligence model first uses the webcam to record a hand sign, classify the word associated with the sign, and convert the audio A similar vein, the suggested system can use Python sound and the OpenCV framework to translate voice to sign. This is used to build the twoway communication approach.

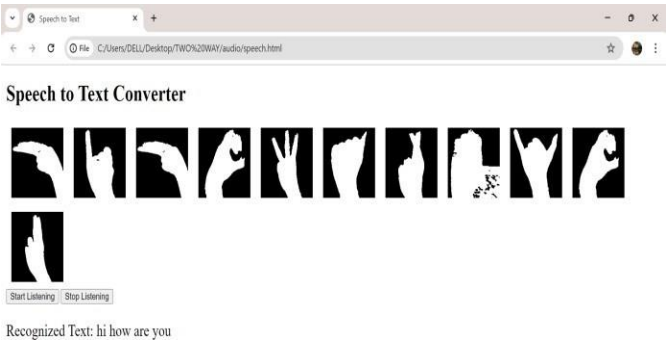


Fig.2 Voice to Gesture output

According to the abstract, the suggested system presents a cutting-edge method for Speech Emotion Recognition (SER) that blends deep learning methods with conventional audio analysis. Mel-frequency cepstral coefficients (MFCC), a tried-and-true technique in speech processing, are first extracted from audio samples as the core feature set. This method is unique in that it turns these MFCC features into MFCC images afterwards. Convolutional Neural Networks (CNNs), which are generally employed for computer vision tasks, can now be applied to the task of emotion categorization thanks to this change. Specifically designed to interact with these MFCC images, the CNN

architecture makes use of its spatial pattern recognition skills to identify emotional cues buried in the data.

With its ability to bridge the gap between audio and visual data representations, this novel SER technique has great potential for speech recognition of subtle emotional expressions. In order to improve the accuracy of emotion recognition, it tackles the requirement for feature-rich representations and sophisticated modelling techniques.

A combination of CNN and MFCC opens up new possibilities for practical uses, such as sentiment analysis, mental health monitoring, and virtual assistants, where it's critical to comprehend speech-based human emotions. Subsequent efforts will concentrate on refining and validating this groundbreaking SER methodology by increasing the dataset and optimizing the bespoke CNN architecture.

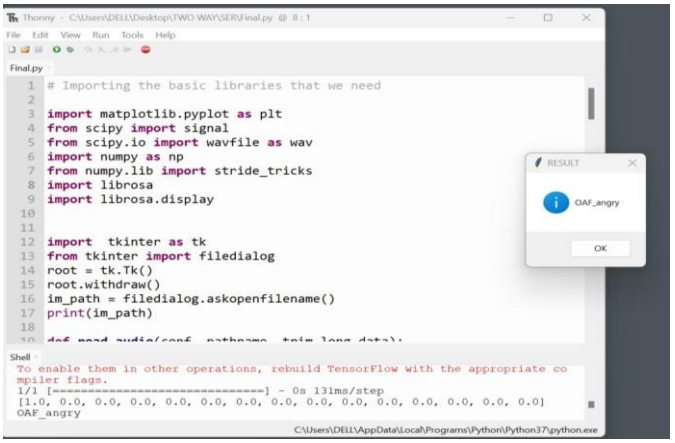


Fig.3 SER output

5. CONCLUSION:

The suggested system, which may be implemented on any processor device, incorporates a two-way sign language translator to close the communication gap between the hearing and the deaf communities. Because it can process an RGB input from a standard camera, the DNN architecture used for sign to text translation is particularly well suited for this application. However, this suggested effort is a very efficient way to translate sign language to text faster.

In conclusion, the abstract's suggested MFCC-based CNN method for Speech Emotion Recognition (SER) offers a creative and promising path for emotional intelligence applications and human-computer interaction. This approach, which unites deep learning with conventional audio analysis approaches, uses Mel-frequency cepstral coefficients (MFCC) to extract features and converts them into MFCC pictures, which Convolutional Neural Networks (CNNs) may use. The possibility to capture subtle emotional expressions in speech is presented by this combination of visual and auditory data representation, addressing the need for more sophisticated modelling techniques and richer feature sets in speech recognition (SER). Thus, the suggested method has a lot of potential for practical uses such as sentiment analysis, mental health monitoring, and virtual assistants, where it is crucial to recognize and react to human emotions expressed through speech. This novel SER methodology is expected to be further refined and validated by future work in increasing the dataset and refining the tailored CNN architecture, potentially yielding significant increases in emotion detection accuracy and usefulness in a range of real-world scenarios.

6. REFERENCES:

[1] S.H. Lee, M.K. Sohn, D.J. Kim, B. Kim, and H. Kim, "Smart TV interaction system using face and hand gesture recognition," in Proc. ICCE, Las Vegas, NV, 2013, pp. 173-174.

[2] S. Kim, G. Park, S. Yim, S. Choi and S. Choi, "Gesture-recognizing hand-held interface with vibrotactile feedback for 3D interaction," IEEE Trans. Consum. Electron., vol. 55, no. 3, pp. 1169-1177, 2009.

[3] S. S. Rautaray, and A. Agrawal, "Vision based hand gesture recognition or human computer interaction: a survey," Artificial Intelligence Review, vol. 43, no. 1, pp. 1-54, 2015.

[4] D. W. Lee, J. M. Lim, J. Sunwoo, I. Y. Cho and C. H. Lee, "Actual remote control: a universal remote control using hand motions on a virtual menu," IEEE Trans. Consum. Electron., vol. 55, no. 3, pp. 1439-1446, 2009.

[5] D. Lee and Y. Park, "Vision-based remote control system by motion detection and open finger counting," IEEE Trans. Consum. Electron., vol. 55, no. 4, pp. 2308-2313, 2009.

[6] F. Erden and A. E. Çetin, "Hand gesture based remote control system using infrared sensors and a camera," IEEE Trans. Consum. Electron., vol. 60, no. 4, pp. 675-680, 2014.

[7] S. Jeong, J. Jin, T. Song, K. Kwon and J. W. Jeon, "Single-camera dedicated television control system using gesture drawing," IEEE Trans. Consum. Electron., vol. 58, no. 4, pp. 1129-1137, 2012

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE Trans. on Pattern Anal. Mach. Intell., vol. 38, no. 1, pp. 142-158,2016.

[9] R. Girshick, "Fast R-CNN," in Proc. ICCV, Santiago, 2015, pp. 1440-1448.

[10] S. Ren, K. he, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans. on Pattern Anal. Mach. Intell., vol.PP, no.99, pp.1-1, 2016.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. CVPR, Las Vegas, NV, 2016, pp. 779-788.

[12] P. Molchanov, S. Gupta, K. Kim, and J. Kautz "Hand gesture recognition with 3D convolutional neural networks," in Proc. CVPR, Boston, MA, 22015, pp. 1-7.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in Proc. ECCV, Zurich,
2014, pp. 346-361.

[14] M. A. Nielsen (2015, January 1). Neural Networks and Deep Learning,.

[15] S. J. Nowlan, and G. E. Hinton. "Simplifying neural networks by soft weight-sharing," Neural computation, vol. 4, no. 4, pp. 473-493, 1992.

[16] G. E. Hinton, Geoffrey, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. (2012, July) Improving neural networks by preventing co-adaptation of feature detectors. Cornell University Library, NY. [Online]. Available: https://arxiv.org/pdf/1207.0580.pdf.

[17] N. H. Dardas, and N. D. Georganas. "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques," IEEE Trans. on Instrum. Meas., vol. 60, no. 11, pp. 3592-3607, 2011.

[18] X. Liu, and F. Kikuo, "Hand gesture recognition using depth data," in Proc. ICAFGR, 2004, pp. 529-534.

[19] S. B. Wang, A. Quattonni, L. P. Morency, and D. Demirdjian, "Hidden conditional random fields for gesture recognition," in Proc. ICCV, 2006, pp. 1521-1527.

[20] P. Trindade, J. Lobo, and J. P. Barreto, "Hand gesture recognition using color and depth images enhanced with hand angular pose data," in Proc. ICMFIIS, Hamburg, 2012, pp. 71-76.

[21] A. I. Maqueda, C. R. del-Blanco, F. Jaureguizar, and N. García, "Human– computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns," Comp. Vis. Image Underst., vol 141, pp.

[22] Kotikalapudi Vamsi Krishna,Navuluri Sainath,A. Mary Posonia,"Speech Emotion Recognition using Machine Learning",2022 6th International Conference on Computing Methodologies and Communication (ICCMC)

[23] R. Anusha,P. Subhashini,Darelli Jyothi,Potturi Harshitha,Janumpally Sushma,Namsamgari Mukesh,"Speech Emotion Recognition using Machine Learning",2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)

[24] Kartikeya Srinivas Chintalapudi,Irfan Ali Khan Patan,Harsha Vardhan Sontineni,Venkata Saroj Kushwanth Muvvala,SuryaKanth V Gangashetty,Akhilesh Kumar Dubey,"Speech Emotion Recognition Using Deep Learning",2023 International Conference on Computer Communication and Informatics (ICCCI)

[25] S. Lalitha,Abhishek Madhavan,Bharath Bhushan,Srinivas Saketh,"Speech emotion recognition",2014 International Conference on Advances in Electronics Computers and Communications

[26]] Girija Deshmukh,Apurva Gaonkar,Gauri Golwalkar,Sukanya Kulkarni,"Speech based Emotion Recognition using Machine Learning",2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)

[27] Chen Jie,"Speech emotion recognition based on convolutional neural network",2021 International Conference on Networking Communications and Information Technology (NetCIT)