



ISSN: 2454-9940



**INTERNATIONAL JOURNAL OF APPLIED
SCIENCE ENGINEERING AND MANAGEMENT**

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

Face And Audio based Emotion Detection using Deep Learning

M Anil Kumar¹, Dr U M Fernandes Dimlo²

¹ PG, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India.

² Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India.

ABSTRACT

Artificial intelligence, machine learning, human-machine interface, etc., have all seen consistent innovation over the last several years. The use of voice commands to direct machines to do specific actions is rapidly expanding. There are a plethora of built-in consumer devices, including Alexa, SIRI, Cortana, Google Assistant, and the like. Unfortunately, computers can't have a genuine conversation like a person. It can't read people's feelings or react to them. Research on emotion identification from speech is considered cutting edge in the area of human-computer interaction. We must strengthen our system of human-machine communication since manufacturers are crucial to our existence. Scientists are now interested in speech emotion recognition (SER) as a means to enhance human-machine communication. If we want a computer to do this, it has to be able to detect when people are feeling sad or angry and respond appropriately. The quality of the features gathered and the kind of classifiers used determine how well the speech emotion recognition (SER) system performs. Using just verbal cues, we set out to determine whether subjects were furious, dissatisfied, neutral, or delighted. Here, audio clips of brief Manipuri speech taken from films served as the training and screening datasets. Here, you'll use the Mel Regularity Cepstral Coefficient (MFCC) function extraction method to train a CNN to identify various emotions in spoken English.

Introduction

Acknowledging the face, removing and classifying features, and using voice sounds to convey thoughts and feedback are the three steps involved in automated facial emotion recognition. Despite significant progress in human user interface technologies, including as the ubiquitous mouse and keyboard, automatic voice recognition, and accessible interfaces for people with impairments, these essential interactive skills are often disregarded. Consequently, consumers often encounter inadequate services. A better tailored experience that met the user's needs and exceeded their expectations would be possible if computers could detect these emotional signals. The six archetypal human emotions identified by psychological study are shock, terror, disgust, frenzy, pleasure, and suffering. Nonverbal clues, such voice intonation and facial expression, play a significant role in conveying emotional states.

An exciting new field of study, emotion interpretation may provide answers to many mysteries. Body language and facial expressions are ways individuals communicate themselves emotionally, whether on purpose or by accident. Vocal, creative, and aesthetic data are only a few of the many forms of information that may be used for emotion analysis. The most reliable ways to read people's emotions, which include their thoughts, have long been their speech and facial expressions. Discovering the underlying emotions and sentiments is a daunting and difficult undertaking. In response, researchers from a wide range of fields are focusing on improving methods for detecting emotional states in various electrical signals, including those emitted by the voice and the face.

Artificial intelligence, natural language modelling systems, etc., have been used to enhance the reaction to various vocal-based techniques and speeches. Numerous fields might benefit from emotional analysis. An example of this would be working together with human computer systems. Computers can improve decision-making, emotion perception, and the rationality of human-robot interactions. We would find out where emotion data comes from, how to identify emotions, what approaches are now in use for emotion modelling, the pros and cons of these methods, and where future research may go from here. Our primary focus is on studying tasks that involve evaluating emotions via speech and facial recognition. We searched through a plethora of technical libraries that housed modern methods and equipment. We have reached all of the critical market milestones and have discovered strategies that might lead to even better outcomes.

Literature survey:

Recognising Emotions via Speech and Facial Expressions in 2014

Understanding the interplay between people and machines is one of the most dynamic areas in information technology. Asynchronous data from both unimodal and multimodal systems has been the backbone of most research in this area thus far. All of the aforementioned issues with synchronisation contribute to the overall complexity of the system and its impact on response time. In response to this issue, a novel approach has been developed to gauge people's emotional states by analysing their facial expressions and vocal inflections. The method makes use of two feature vectors: the relative bin frequency coefficient (RBFC) for audio data and the family member sub-image based coefficient (RSB) for aesthetic data. Two modalities are combined using a feature-level category based method using a support vector machine with a radial basis bit. The proposed brilliant approach has generated exciting outcomes for a wide variety of inputs, and it is also adaptable to asynchronous data. Relevant keywords include human-computer interaction, AVTs, relative sub-image characteristics, and bin regularity coefficients.

Acknowledgment in 2016 via Facial Expressions and Vocal Expressions in Humans

Abstract-- There has been research on using computer systems to imitate human emotions from the beginning of conversational separation. The goal of this work is to provide a hybrid system that can analyse a person's facial expressions and vocal inflections to determine common human emotions including anger, sadness, joy, boredom, disgust, and astonishment. When it comes to audio data, we use family bin frequency coefficients, and when it comes to visual data, we use loved one sub-image based features. Support Vector Devices trained with radial basis kernels are used in the classification process. The two most crucial parts of an emotion recognition system, according to this study, are the proposed blend approach and function extraction from facial expressions and speech. There are a few factors that could impact the emotion detecting system, but they don't have much of an impact. The bimodal emotion recognition system outperformed the unimodal method, according to calculated faces. Using the right database resolves the problem. The results showed that when competing systems included one of the most basic psychological groupings, the suggested emotion detection system performed better. Search Terms: Support Vector Machines (SVM), Relative Sub-Image Based (RSB), and Dear One Bin Regularity Coefficients (RBFC).

In 2019, Deep Learning Will Power Emotion Recognition.

ABSTRACT Getting a sense of familiarity via vocal indicators is a crucial yet challenging part of HCI. Among the many approaches used in the literature on speech emotion recognition (SER) to demotionalize signals are several well-established speech assessment and classification systems. Using Deep Discovering is a more recent alternative to more traditional approaches in SER. This article covers some recent research that has employed Deep Knowing approaches to identify emotions in spoken language and gives a high-level introduction to these methods. Subjects covered in testimonials include data sources, extracted emotions, acknowledgment of payments to spoken emotions, and related restrictions.

In 2020, we will be able to use CNNs based on facial expressions to acknowledge feelings.

Recent years have seen an explosion in research on facial emotion recognition, mostly because of its practical applications and impact on human-computer communication. As the amount of hard datasets continues to grow, deep learning methods are becoming more important. By examining the challenges of emotion detection datasets and experimenting with different CNN techniques and configurations, we aim to identify the seven human emotions—sourness, anger, fear, disgust, sadness, and shock—in a face scan. Our primary dataset will be the one-of-a-kind, fascinating, and very difficult iCV MEFED (Multi-Emotion Face Dataset). The following keywords are associated with this article: Convolutional Neural Network, Deep Learning, FER, Information Preprocessing, Picture Recognition, Facial Expression Recognition.

Examining the framework:. i.Present configuration:.

Emotion prediction has been a challenge for conversational AI from the start. The authors of this research suggest a hybrid method for gauging important emotions (including surprise, anger, sadness, boredom, contempt, and discontent) from a speaker's words and expressions while they're talking. Relative in regularity coefficients stand in for the auditory data, whereas relative sub-image tures denote the visual data. A Support Vector Machine with a radial basis kernel is used for classification. Feature extraction using facial expressions and speech is the most crucial component of the proposed fusion technique for emotion identification systems, as per this research. There are a few things that can influence the emotion detection system, but they don't have much of an effect. The results demonstrated that the bimodal system achieved better results than the unimodal system when testing with purposeful facial expressions. Using the right database solves the issue. The results showed that for the most basic emotional categories, the proposed emotion recognition system performed better than the competitors.

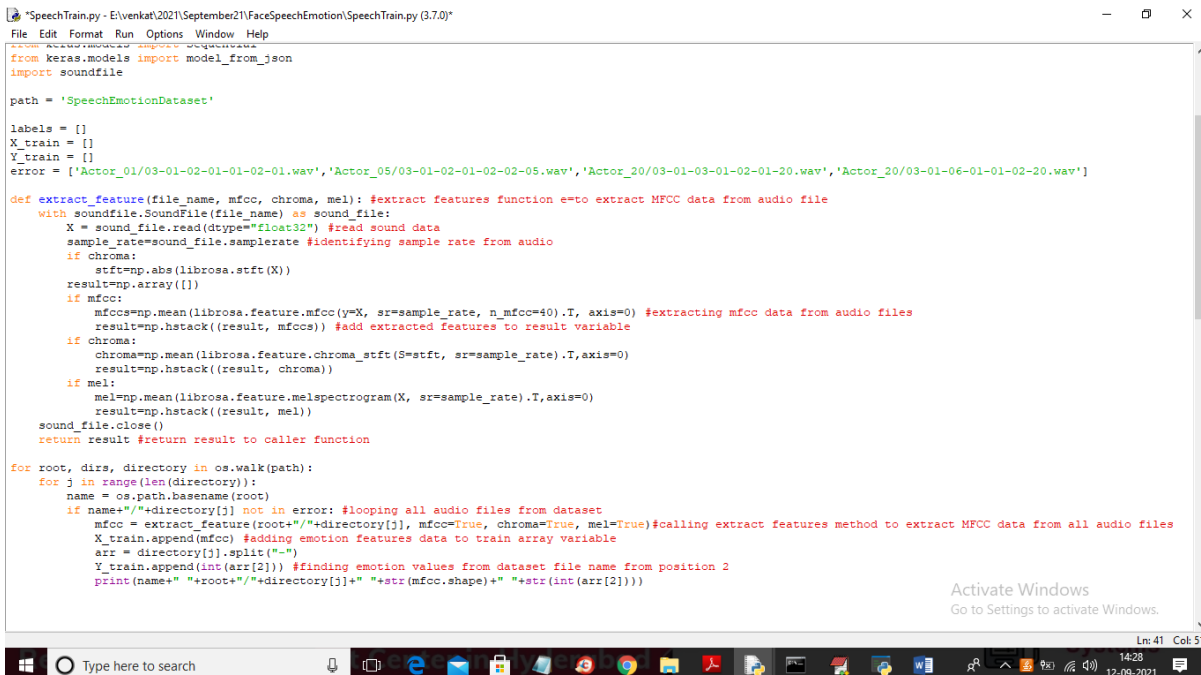
part ii. future setup:.

Recent years have seen a surge in research into facial emotion recognition due to its importance and impact on human-computer interaction. The proliferation of challenging datasets is making deep learning technologies indispensable. Our goal is to recognise the seven human face emotions—angry, fear, disgust, contempt, sadness, surprise, and happiness—by analysing Emotion Recognition Datasets and experimenting with different CNN architectures and settings. We have chosen the innovative, fascinating,

and very challenging iCV MEFED (Multi-Emotion Facial Expression Dataset) as our primary dataset. Issues covered include: facial expression recognition, preprocessing data, CNNs, deep learning, image recognition, and CNNs.

Results:

In order to identify emotions in spoken English, we trained a convolutional neural network (CNN) model using the RAVDESS Audio Dataset and the Emotion Facial Expression images dataset. Extracting MFCC properties from an audio collection is shown in the following code snippets; comments are received in red.



```
SpeechTrain.py - E:\venkat\2021\September21\FaceSpeechEmotion\SpeechTrain.py (3.7.0)
File Edit Format Run Options Window Help
from keras.models import Sequential
from keras.models import model_from_json
import soundfile

path = 'SpeechEmotionDataset'

labels = []
X_train = []
Y_train = []
error = ['Actor_01/03-01-02-01-01-02-01.wav', 'Actor_05/03-01-02-01-02-02-05.wav', 'Actor_20/03-01-03-01-02-01-20.wav', 'Actor_20/03-01-06-01-01-02-20.wav']

def extract_feature(file_name, mfcc, chroma, mel): #extract features function e=to extract MFCC data from audio file
with soundfile.SoundFile(file_name) as sound_file:
X = sound_file.read(dtype="float32") #read sound data
sample_rate=sound_file.samplerate #identifying sample rate from audio
if chroma:
stft=np.abs(librosa.stft(X))
result=np.array([])
if mfcc:
mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0) #extracting mfcc data from audio files
result=np.hstack((result, mfccs)) #add extracted features to result variable
if chroma:
chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
result=np.hstack((result, chroma))
if mel:
mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
result=np.hstack((result, mel))
sound_file.close()
return result #return result to caller function

for root, dirs, directory in os.walk(path):
for j in range(len(directory)):
name = os.path.basename(root)
if name+"/"+directory[j] not in error: #looping all audio files from dataset
mfcc = extract_feature(root+"/"+directory[j], mfcc=True, chroma=True, mel=True)#calling extract features method to extract MFCC data from all audio files
X_train.append(mfcc) #adding emotion features data to train array variable
arr = directory[j].split("-")
Y_train.append(int(arr[2])) #finding emotion values from dataset file name from position 2
print(name+" "+root+"/"+directory[j]+" "+str(mfcc.shape)+" "+str(int(arr[2])))
```

Annotations describing how to extract audio characteristics are shown in red on the top screen. The CNN-trained X and Y data are shown on the bottom screen.

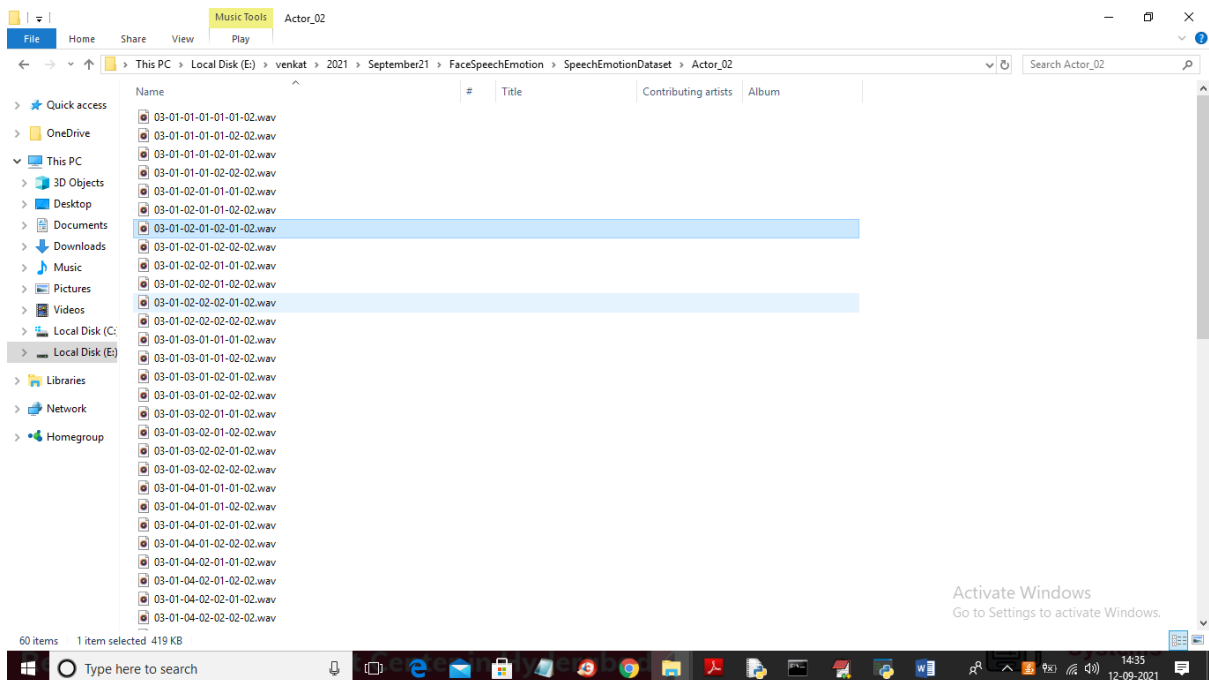

```

SpeechTrain.py - E:\venkat\2021\September21\FaceSpeechEmotion\SpeechTrain.py (3.7.0)
File Edit Format Run Options Window Help
C:\Python37\Scripts\python.exe C:\Python37\Scripts\runpy.py --
with open('model/speechmodel.json', "r") as json_file:
    loaded_model_json = json_file.read()
    classifier = model_from_json(loaded_model_json)
    classifier.load_weights("model/speech_weights.h5")
    classifier.make_predict_function()
    print(classifier.summary())
    f = open('model/speechhistory.pkl', 'rb')
    data = pickle.load(f)
    f.close()
    acc = data['accuracy']
    accuracy = acc[9] * 100
    print("Training Model Accuracy = "+str(accuracy))
else:
    classifier = Sequential() #creating sequential object as classifier
    #creating CNN layer with 32 neurons or filters and giving input shape and this data will be filtered by cnn 32 times
    classifier.add(Convolution2D(32, 1, 1, input_shape = (180, 1, 1), activation = 'relu'))
    #defining max pooling layer to extract important features from dataset
    classifier.add(MaxPooling2D(pool_size = (1, 1)))
    #creating another layer with 32 filters
    classifier.add(Convolution2D(32, 1, 1, activation = 'relu'))
    #defining max pooling layer to extract important features from dataset
    classifier.add(MaxPooling2D(pool_size = (1, 1)))
    #converting multidimensional data to single dimensional data
    classifier.add(Flatten())
    #defining output layer
    classifier.add(Dense(output_dim = 256, activation = 'relu'))
    #output layer has to predict values as per given in Y data
    classifier.add(Dense(output_dim = Y_train.shape[1], activation = 'softmax'))
    #print summary of CNN
    print(classifier.summary())
    #compiling CNN model
    classifier.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
    #start training CNN with given X and Y data
    hist = classifier.fit(X_train, Y_train, batch_size=16, epochs=100, shuffle=True, verbose=2)

    classifier.save_weights('model/speech_weights.h5')
    model_json = classifier.to_json()
    with open('model/speechmodel.json', "w") as jsonFile:
        jsonFile.write(model_json)
    jsonFile.close()
    f = open('model/speechhistory.pkl', 'wb')

```

The algorithm we used to train the convolutional neural network (CNN) on the speech dataset was also used to train the CNN on the face picture dataset, as shown above. The images from the 'Face' dataset that can be viewed on the screen below are stored in the 'Dataset' folder.

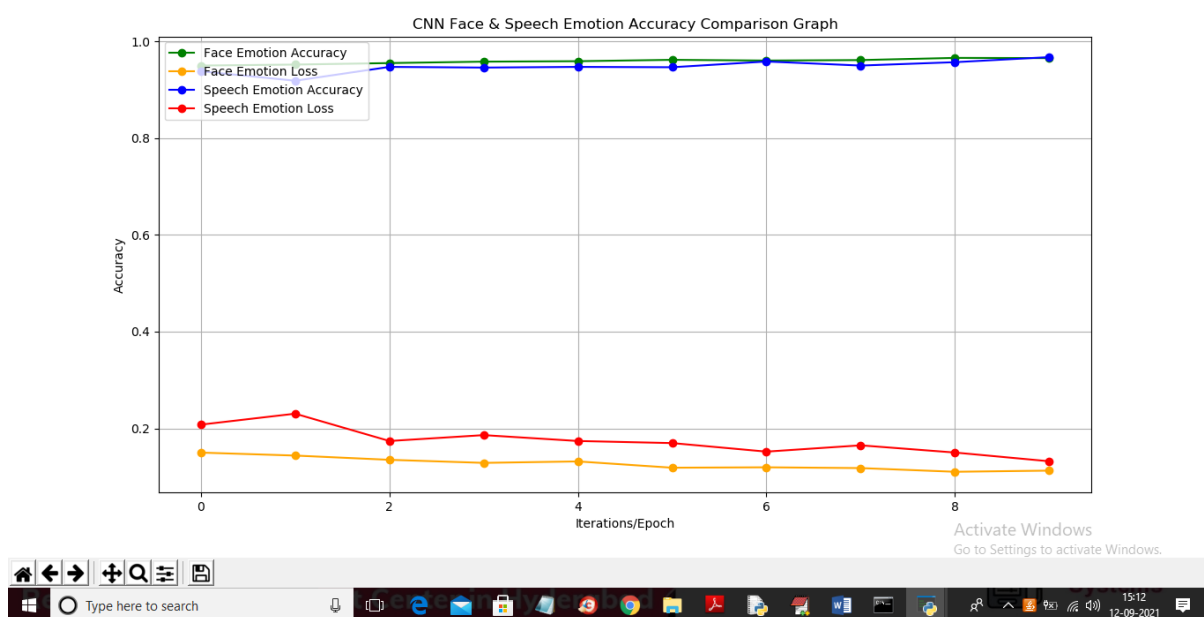


Each wav file is linked to a number separated by a '-' sign; the ID is 03, the gender is 01, and the emotion is the third position value from 1 to 8 on the above screen.

After you double-click the "run.bat" file to start the project, you will see this screen To upload the dataset, click the "Upload Facial Emotion Dataset" button on the previous page. This will bring up the next screen. You may upload a dataset by going to the previous page, finding the "Dataset" folder, and

then clicking the "Select Folder" button. To read all of the photos in the dataset, resize them to the same size, eliminate the MFCC functions, and finally build an experienced model, click the "Preprocess Dataset" button on the top screen. To begin training the Facial dataset using CNN, click the "Train Facial Emotion CNN Algorithm" button; this will bring up the following page, which displays the processed datasets along with the total number of pictures and audio recordings in each. After training the CNN using face photographs, it attained an accuracy of 96.52%; to train it using audio features, click the "Train Speech Emotion CNN Algorithm" button. We achieved an accuracy of 96.72% using CNN Speech Emotion in the above screen. To see the graph below, click on the "Accuracy Comparison Graph" button.

Figure 1



Both algorithms' loss values dropped to zero, and the x-axis shows the date, while the y-axis shows the accuracy and loss values. In the graph above, the blue line reflects accurate speech, whereas the green line shows accurate facial expressions. To get the results shown below, upload a picture of your face and then click the "Predict Facial Emotion" button.

After choosing and uploading the '5.jpg' picture in the previous screen, click on the 'Open' button to get the following outcome:

We may now try out other images to see whether they match the "Fearful" predicted expression on the previous screen.



You may add more photos and rate them similarly; the main page will provide you a "happy" mood prediction. After you've uploaded the audio recording, click the "Predict Speech Emotion" button to get the result that you see below.

Here is the outcome of choosing and uploading the "2.wav" file on the previous screen 'Calm' is the anticipated emotion for the audio track you submitted in the previous screen; now try out another file.

You can see the outcome of the prediction in the preceding screen after uploading the "5.wav" file.



The uploaded file is shown with an emotion prediction of "angry" on the above screen; you may upload and test more files in a similar fashion.

Conclusion:

A new dual-channel expression recognition algorithm based on AI idea and emotional perspective is proposed in this study. The first step of the suggested algorithm uses the Gabor attribute of the ROI region as input as functions derived using CNNs miss fine-grained changes in the expressive regions of the face. In order to fully use the detail feature of the active face region, the initial face picture is used to segment the active face area. Then, the characteristics of this area are extracted using Gabor change, with a greater emphasis on the detail summary of the local region. A channel focus network built upon deep separable convolution is presented in the second route to minimise overfitting, reduce network complexity, and enhance the straight bottleneck framework. By focusing more on the extraction of critical characteristics and boosting the accuracy of emotion detection, an effective interest module is designed to include spatial information into the depth of the attribute map. Competitive efficiency was achieved on the FER2013 data sets.

Furthermore, this research will serve as a manual for advancing people's agency and proves that Zeng Guofan's perspective on human resource management is sound.

REFERENCES:

- [1] Bjorn S, Stefan S, Anton B, Alessandro V, Klaus S, Fabien R, Mohamed C, Felix W, Florian E, Erik M, Marcello M, Hugues S, Anna P, Fabio V, Samuel K (2013) Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism
- [2] Deepak G, Joonwhoo L (2013) Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. *Sensors* 13:7714–7734. <https://doi.org/10.3390/s130607714>
- [3] Domínguez-Jiménez JA, Campo-Landines KC, Martínez-Santos J, Delahoz EJ, Contreras-Ortiz S (2020) A machine learning model for emotion recognition from physiological signals. *Biomed Signal Proces* 55:101646
- [4] El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn* 44:572–587. <https://doi.org/10.1016/j.patcog.2010.09.020>
- [5] Eyben F, Scherer KR, Schuller BW et al (2016) The Geneva minimalistic acoustic parameter set (gaMAPS) for voice research and affective computing. *IEEE Trans Affect Comput* 7:190–202. <https://doi.org/10.1109/TAFFC.2015.2457417>
- [6] Ghimire D, Jeong S, Lee J, Park SH (2017) Facial expression recognition based on local region specific features and support vector machines. *Multimed Tools Appl* 76:7803–7821. <https://doi.org/10.1007/s11042-016-3418-y>
- [7] Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. <https://www.deeplearningbook.org>. Accessed 1 Mar 2020
- [8] Hamm J, Kohler CG, Gur RC, Verma R (2011) Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *J Neurosci Methods* 200:237–256
- [9] Happy SL, George A, Routray A (2012) A real time facial expression classification system using local binary patterns. In *Proc 4th Int Conf Intell Human Comput Interact* 27–29:1–5
- [10] Hasani B, Mahoor MH (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks. *IEEE Conf Comput Vision Pattern Recognit Workshops (CVPRW)*. <https://doi.org/10.1109/CVPRW.2017.282>

- Li D, Bo S, Yu L (2019) Facial action unit detection with multilayer fused multi-task and multi-label deep learning network. *KSII Trans Internet Inf Syst* 7:5546–5559. <https://doi.org/10.3837/tis.2019.11.015>
- [1] Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio-visual emotional big data. *Inf Fusion* 49:69–78. <https://doi.org/10.1016/j.inffus.2018.09.008>
- [2] Hutto CJ, Eric G (2014) VADER: A parsimonious rule-based model for sentiment analysis of social media text. *AAAI Publications, Eighth Int AAAI Conf Weblogs Soc Media*
- [3] Iliou T, Anagnostopoulos C-N (2009) Statistical evaluation of speech features for emotion recognition. In: *Digital telecommunications ICDT'09 4th Int Conf IEEE* 121–126
- [4] Jia X, Li W, Wang Y, Hong S, Su X (2020) An action unit co-occurrence constraint 3DCNN based action unit recognition approach. *KSII Trans Internet Inf Syst* 14:924–942. <https://doi.org/10.3837/tis.2020.03.001>
- [5] Joseph R, Santosh D, Ross G, Ali F (2015) You Only Look Once: Unified, Real-Time Object Detection arXiv preprint arXiv:1506.02640
- [6] Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. *2015 IEEE Int Conf Comput Vision (ICCV)* <https://doi.org/10.1109/ICCV.2015.341>
- [7] Kao YH, Lee LS (2006) Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: *InterSpeech*
- [8] Kaulard K, Cunningham DW, Bülthoff HH, Wallraven C (2012) The MPI facial expression database—A validated database of emotional and conversational facial expressions. *PLoS One* 7:e32321.
- [9] Khan RA, Meyer A, Konik H, Bouakaz S (2013) Framework for reliable, real-time facial expression recognition for low resolution images. *Pattern Recogn Lett* 34:1159–1168. <https://doi.org/10.1016/j.patrec.2013.03.022>
- [10] Ko BC (2018) A brief review of facial emotion recognition based on visual information. *Sensors* 18. <https://doi.org/10.3390/s18020401>
- [11] LeCun Y, Bengio Y, Hinton G (2015) Deep learning, *Nature* 521. <https://doi.org/10.1038/nature14539>

Lee C, Lui S, So C (2014) Visualization of time-varying joint development of pitch and dynamics for speech emotion recognition. *J Acoust Soc Am* 135:2422. <https://doi.org/10.1121/1.4878044>

- [1] Li S, Deng W (2020) Deep facial expression recognition: A survey. *IEEE Trans Affective Comp (Early Access)*. <https://doi.org/10.1109/TAFFC.2020.2981446>
- [2] Liu M, Li S, Shan S, Wang R, and Chen X (2014) Deeply learning deformable facial action parts model for dynamic expression analysis. 2014 Asian Conference on Computer Vision (ACCV) 143–157. https://doi.org/10.1007/978-3-319-16817-3_10
- [3] Lotfian R, Busso C (2019) Curriculum learning for speech emotion recognition from crowdsourced labels. *IEEE/ACM Trans Audio, Speech Lang Processing* 4. <https://doi.org/10.1109/TASLP.2019.2898816>
- [4] Luengo I, Navas E, Hernandez I, Sanchez J (2005) Automatic emotion recognition using prosodic parameters. In: *Interspeech*, 493–496
- [5] Ma Y, Hao Y, Chen M, Chen J, Lu P, Koar A (2019) Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Inf Fusion* 46:184–192. <https://doi.org/10.1016/j.inffus.2018.06.003>
- [6] Mehrabian A (1968) Communication without words. *Psychol Today* 2:53–56
- [7] Mira J, ByoungChul K, JaeYeal N (2016) Facial landmark detection based on an ensemble of local weighted regressors during real driving situation. *Int Conf Pattern Recognit* 1–6.
- [8] Mira J, ByoungChul K, Sooyeong K, JaeYeal N (2018) Driver facial landmark detection in real driving situations. *IEEE Trans Circuits Syst Video Technol* 28:2753–2767. <https://doi.org/10.1109/TCSVT.2017.2769096>
- [9] Rao KS, Koolagudi SG, Vempada RR (2013) Emotion recognition from speech using global and local prosodic features. *Int J Speech Technol* 16(2):143–160
- [10] Scherer KR (2003) Vocal communication of emotion: A review of research paradigms. *Speech Comm* 40:227–256. [https://doi.org/10.1016/S0167-6369\(02\)00084-5](https://doi.org/10.1016/S0167-6369(02)00084-5).
- [11] Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Comm* 53(9–10):1062–1087.
- [12] Shaqr FA, Duwairi R, Al-Ayyou M (2019) Recognizing emotion from speech based on age and gender using hierarchical models. *Procedia Comput Sci* 151:37–44. <https://doi.org/10.1016/j.procs.2019.04.009>

Si

- ddiqi MH, Ali R, Khan AM, Park YT, Lee S (2015) Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. *IEEE Trans Image Proc* 24:1386–1398. <https://doi.org/10.1109/TIP.2015.2406346>
- [1] Song P, Zheng W (2018) Feature selection based transfer subspace learning for speech emotion recognition. *IEEE Trans Affective Comput (Early Access)* <https://doi.org/10.1109/TAFFC.2018.2800046>
- [2] Sun N, Qi L, Huan R, Liu J, Han G (2019) Deep spatial-temporal feature fusion for facial expression recognition in static images. *Pattern Recogn Lett* 119:49–61. <https://doi.org/10.1016/j.patrec.2017.10.022>
- [3] Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speech emotion recognition: A review. *Int J Speech Technol* 21:93–120. <https://doi.org/10.1007/s10772-018-9491-z>
- [4] Wang X, Chen X, Cao C (2020) Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Process Image Commun* <https://doi.org/10.1016/j.image.2020.115831>
- [5] Wu CH, Yeh JF, Chuang ZJ (2009) Emotion perception and recognition from speech. *Affective Inf Processing* 93–110. https://doi.org/10.1007/978-1-84800-306-4_6
- [6] Xiong X and Fernando DIT (2013) Supervised descent method and its applications to face alignment. 2013 *IEEE Conf Comput Vision and Pattern Recognit (CVPR)* <https://doi.org/10.1109/CVPR.2013.75>
- [7] Zamil AAA, Hasan S, Baki SJ, Adam J, Zaman I (2019) Emotion detection from speech signals using voting mechanism on classified frames. 2019 *Int Conf Robotics, Electr Signal Processing Technol (ICREST)* <https://doi.org/10.1109/ICREST.2019.8644168>
- [8] Zhang H, Huang B, Tian G (2020) Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recogn Lett* 131:128–134. <https://doi.org/10.1016/j.patrec.2019.12.013>
- [9] Zhang S, Zhang S, Huang T, Gao W (2008) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans Multimed* 20:1576–1590. <https://doi.org/10.1109/TMM.2017.2766843>
- [10] Zhang T, Zheng W, Cui Z, Zong Y, Yan J, Yan K (2016) A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Trans Multimed* 18:2528–2536. <https://doi.org/10.1109/TMM.2016.2598092>