# Emotion recognition using sound

## Methods

1-Feature Extraction

used to  two recent advanced algorithms or **multilayers deep learning paradigms**

- Wav2vec2.0

  Wav2vec2.0 is a self-supervised speech representation model that pursues to capture the crucial properties of raw audios by using the power of transformers and Contrastive learning.

  consists of

  -convolutional layers that process the raw waveform input to get latent representation – Z.

  - transformer layers, creating contextualized representation – C linear projection to output – Y.

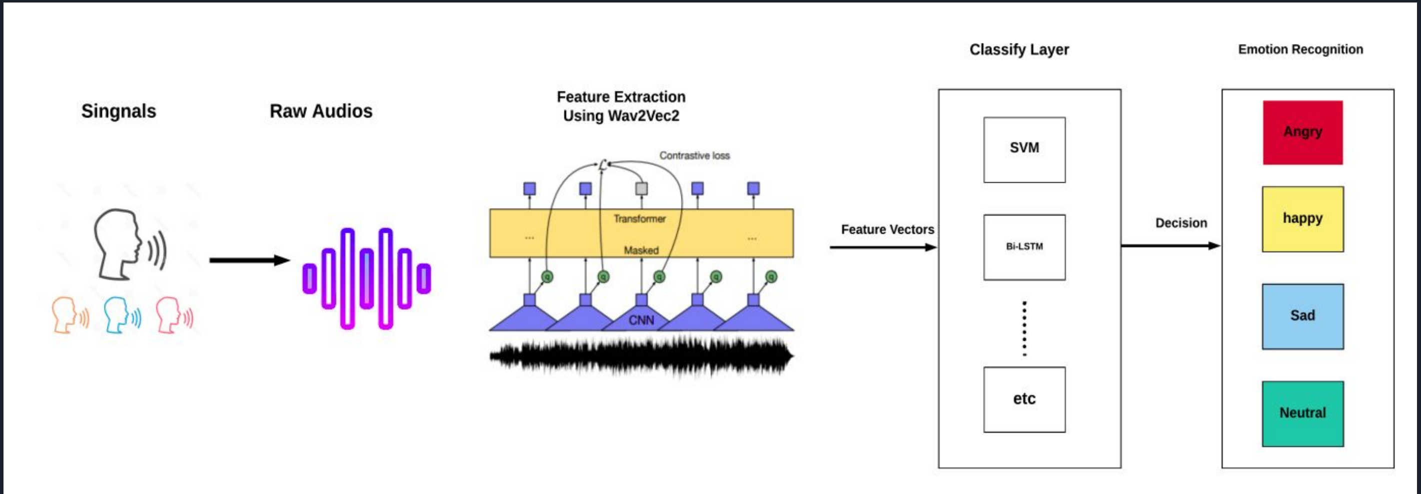
- HuBERT

  Innovative method for self-supervised speech representation learning HuBERT for speech representation learning matches or outperforms SOTA techniques for speech recognition, generation, and compression.

  HuBERT learns the structure of spoken input by predicting the proper cluster for masked audio segments using an offline k-means clustering step. By alternating between clustering and prediction processes uses continuous inputs to train both acoustic and linguistic models. The model must first encode unmasked audio inputs into meaningful continuous latent representations, which correspond to the traditional acoustic modelling problem. Second, the model must capture the long-term temporal relationships between learned representations in order to reduce prediction error.

## 2-MLP and Bi-LSTM Classifiers

After extracting features with wav2vec2.0 and HuBERT, we feed the output into a classifier head: We utilized MLP Classifier stands for Multi-layer Perception Classifier, which is linked to a Neural Network by its name, and a Bi-LSTM Layer with 50 hidden units, Both classifiers produced results that were close to each other.



| model | Length | no. records | accuracy |
|---|---|---|---|
| wav2vec2.0 | 19 Min | 1935 | 89 |
| Hubert Base | 19 Min | 1935 | 87 |
| Hubert Large | 19 Min | 1935 | 84 |

-----------------------------------------------------------------------------------------------------------

## ARCHITECTURE DESIGN

1-Input pipeline

After normalizing the audio signals between −1 and 1, the MFCCs of the signals are calculated. we use a Hamming window to split the audio signal into 64-ms frames with 16ms overlaps, which can be considered as quasi-stationary segments. Following a 1024-point Fast Fourier transform (FFT) applied to each frame, the signal undergoes a Mel scale filter bank analysis.

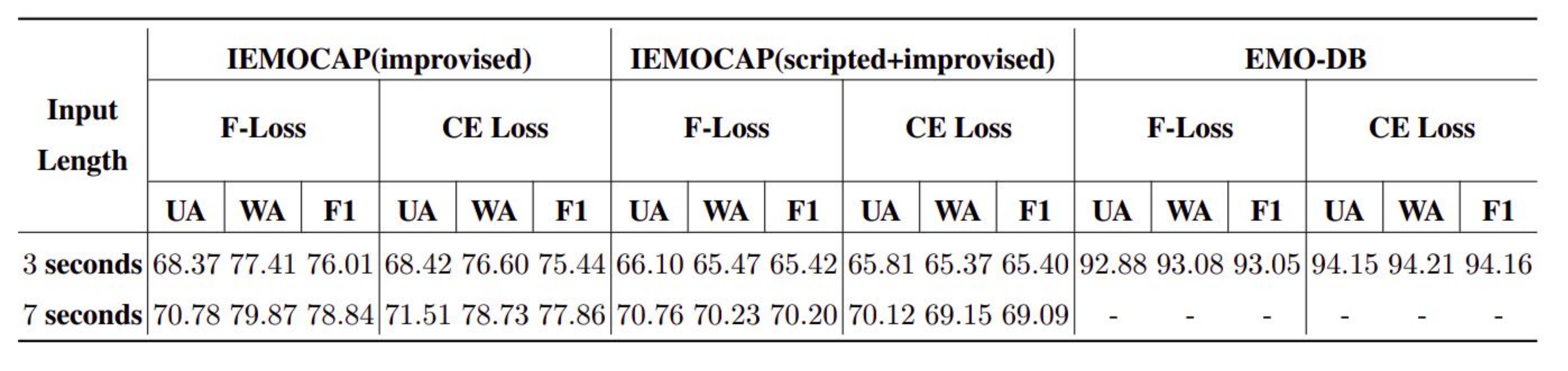
2-Body Part I

three parallel CNNs are applied to the MFCC to extract time and frequency features. This structure can achieve a balance between spectral and temporal information in its feature extractor.

3-Body Part II

consists of several LFLBs with different configurations applied to the concatenated low-level features from Body part I to capture high-level features

4- Head The body part is supposed to map the nonlinear input space into a linearly separable sub-space, and thus, one fullyconnected layer is enough for the classification.



| Input Length | IEMOCAP(improvised) | | | | | | IEMOCAP(scripted+improvised) | | | | | | EMO-DB | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F-Loss | | | CE Loss | | | F-Loss | | | CE Loss | | | F-Loss | | | CE Loss | | |
| | UA | WA | F1 | UA | WA | F1 | UA | WA | F1 | UA | WA | F1 | UA | WA | F1 | UA | WA | F1 |
| 3 seconds | 68.37 | 77.41 | 76.01 | 68.42 | 76.60 | 75.44 | 66.10 | 65.47 | 65.42 | 65.81 | 65.37 | 65.40 | 92.88 | 93.08 | 93.05 | 94.15 | 94.21 | 94.16 |
| 7 seconds | 70.78 | 79.87 | 78.84 | 71.51 | 78.73 | 77.86 | 70.76 | 70.23 | 70.20 | 70.12 | 69.15 | 69.09 | - | - | - | - | - | - |

# Multimodal Emotion Recognition using Deep Learning

## Methods

### 1- Facial expression recognition

Mehrabian observed that 7% of knowledge moves between people through writing, 38% through voice, and 55% through facial expression [36]. Ekman et al. [37] defined six basic emotions: happiness, sadness, surprise, fear, and anger.

### 2-Speech emotion recognition

The earliest experiments on emotion recognition in speech considered extraction of handcrafted features of speech for classification. Many people proposed machine learning algorithms like Support vector machines, hidden Markov models, Gaussian mixture models, etc. Deep learning has been widely used in several speech domains, including speech recognition . Convolution neural network has also been used for speech emotion recognition. shows that using RNN bidirectional- (Bi-LSTM) is better for extracting essential speech characteristics for better speech recognition performance .

## 3-Multimodal emotion recognition

A. Multimodal emotion recognition combining (audio and text, image and text)

-hybrid fusion process, referred to as a multimodal attention network (MMAN), to make use of visual and textual signals in speech emotion recognition. They suggest a new multimodal focus -mechanism, cLSTM-MMA, which promotes attention across three modalities and fuses information selectively. During late fusion, cLSTM-MMA is fused with other uni-modal subnetworks. The tests demonstrate that identifying speech emotions profits immensely from visual and textual signals. The suggested cLSTM-MMA alone is as successful in terms of precision as other fusion approaches but with a much more compact network structure.

-searched the use of the pretrained "BERT-like" architecture for self-supervised learning (SSL) to both represent language and text modalities in order to recognize the multimodal language emotions

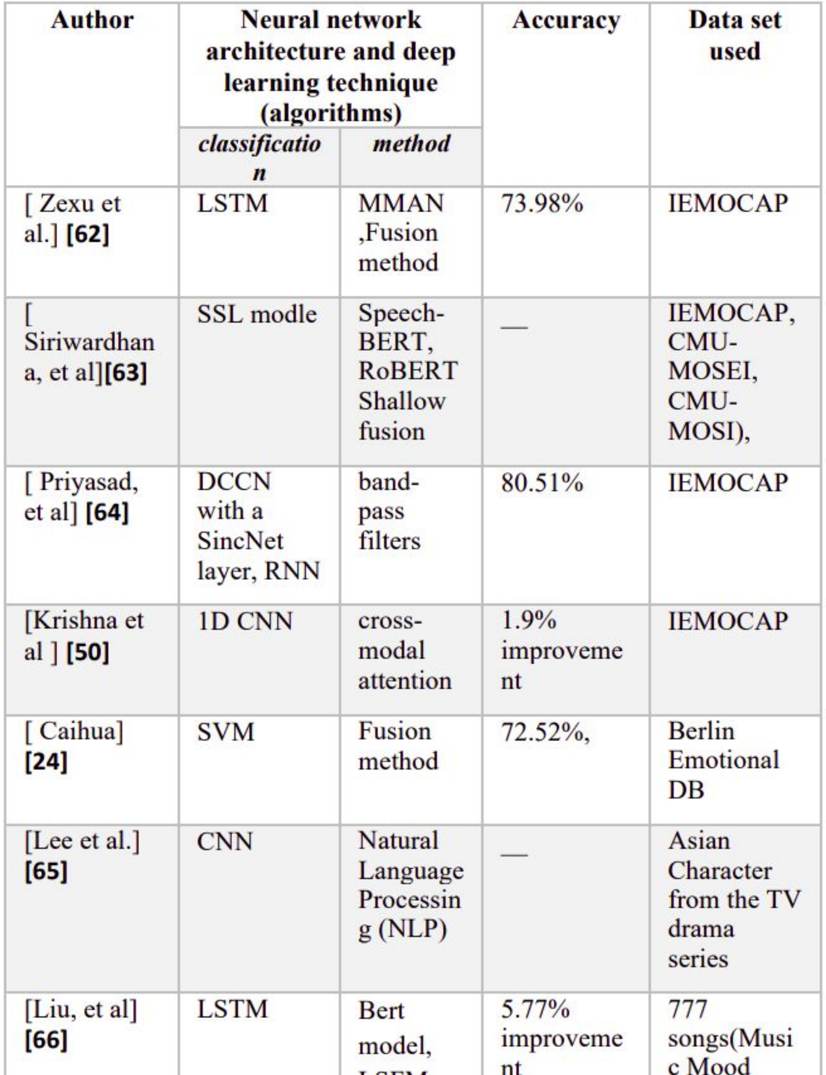
- protect codes that are characteristic emotion. Through a SincNet layer, band-pass filtering technique and neural net,features from raw audio, and the output of said band-pass filters is then applied to the input to DCNN

-a raw-waveform-based convolutional neural network with cross-modal focus. They use raw audio processing by using one-dimensional convolutional models and attention processes between the audio and text feature to obtain the enhanced emotion detection

-through tests, he found that the SVM-based approach of machine learning is powerful for voice consumer sentiment analysis. He suggested an SVM-based multimodal speech emotion recognition

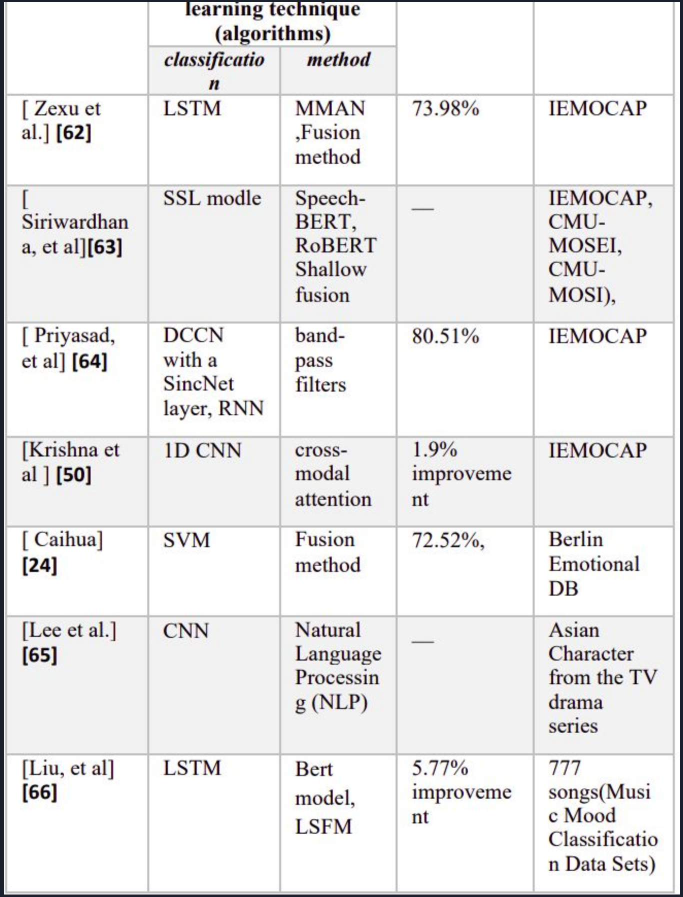
-developed a multimodal deep learning model that utilizes facial images and textual details explaining the circumstances. To classify the characters' facial images in the Korean TV series 'Misaeng

-A new multimodal music emotion grouping system was developed based on music audio quality and text for music lyrics. Use of the LSTM network for classification is suggested in terms of audio, and the classification effect is greatly increased relative to other machine learning methods



| Author | Neural network architecture and deep learning technique (algorithms) | | Accuracy | Data set used |
|---|---|---|---|---|
| | *classification* | *method* | | |
| [ Zexu et al.] **[62]** | LSTM | MMAN ,Fusion method | 73.98% | IEMOCAP |
| [ Siriwardhana, et al]**[63]** | SSL modle | Speech-BERT, RoBERT Shallow fusion | — | IEMOCAP, CMU-MOSEI, CMU-MOSI), |
| [ Priyasad, et al] **[64]** | DCCN with a SincNet layer, RNN | band-pass filters | 80.51% | IEMOCAP |
| [Krishna et al ] **[50]** | 1D CNN | cross-modal attention | 1.9% improvement | IEMOCAP |
| [ Caihua] **[24]** | SVM | Fusion method | 72.52%, | Berlin Emotional DB |
| [Lee et al.] **[65]** | CNN | Natural Language Processing (NLP) | — | Asian Character from the TV drama series |
| [Liu, et al] **[66]** | LSTM | Bert model, LSTM | 5.77% improvement | 777 songs(Music Mood |

# Multimodal emotion recognition combining (facial and body physiological )



| | learning technique (algorithms) | | | |
|---|---|---|---|---|
| | ***classification*** | ***method*** | | |
| [ Zexu et al.] **[62]** | LSTM | MMAN ,Fusion method | 73.98% | IEMOCAP |
| [ Siriwardhana, et al]**[63]** | SSL modle | Speech-BERT, RoBERT Shallow fusion | — | IEMOCAP, CMU-MOSEI, CMU-MOSI), |
| [ Priyasad, et al] **[64]** | DCCN with a SincNet layer, RNN | band-pass filters | 80.51% | IEMOCAP |
| [Krishna et al ] **[50]** | 1D CNN | cross-modal attention | 1.9% improvement | IEMOCAP |
| [ Caihua] **[24]** | SVM | Fusion method | 72.52%, | Berlin Emotional DB |
| [Lee et al.] **[65]** | CNN | Natural Language Processing (NLP) | — | Asian Character from the TV drama series |
| [Liu, et al] **[66]** | LSTM | Bert model, LSFM | 5.77% improvement | 777 songs(Music Mood Classification Data Sets) |