1-Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional Neural Network 2021
Link:
https://link.springer.com/chapter/10.1007/978-981-16-4625-6_24

2-Emotion Recognition from Speech Signals Using Machine Learning and Deep Learning Techniques 2021
Link:
https://link.springer.com/chapter/10.1007/978-3-030-76167-7_4

3-Human emotion recognition by optimally fusing facial expression and speech feature 2020.
Link:
https://www.sciencedirect.com/science/article/abs/pii/S0923596520300540

1-Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional Neural Network

- This paper presents a three-layered sequential deep convolutional neural network (DCNN) based on mel frequency log spectrogram (MFLS) for emotion recognition.

- Dataset:e Berlin Emo-DB

- Accurency: 95.68% for the speaker-dependent approaches
  96.07% for speaker-independent approaches

- The performance of the proposed method is compared with CNN and CNN-LSTM on the Berlin Emo-DB dataset and results in improved accuracy

Human speech signal

- consists of verbal and para verbal information.
  - The verbal information → describes the meaning and context of the speech
  - the para verbal information → describes the tacit information such as the emotion expressed in the speech signal.

- Different emotion has an immense effect on the various characteristics of the speech signal
  - short-term features like energy, pitch and format
  - long-term features like mean and standard deviation
  - prosodic features like pitch, speaking rate, intensity, voice variation and quality

- There are three types of phonetic features related to human emotional expression:
  - prosodic features.
  - spectral features.
  - phonetic quality features.

Traditional, machine learning (ML)-based SER systems

- Two major phases
  - feature extraction and classification.

- feature extraction techniques for SER system
  - mel frequency cepstrum coefficients (MFCC)
  - principal component analysis (PCA)
  - linear predictor coefficients (LPC)
  - Gaussian mixture model (GMM)
  - perceptual linear prediction coefficients (PLP)
  - hidden Markov model (HMM).

- SER system used various classifications algorithms
  - support vector machine (SVM)
  - K-nearest neighbour classifier (KNN)
  - artificial neural network (ANN).

deep learning (DL) based SER systems

- represents the low-level speech features into the high-level hierarchical features.

- Jianwei Niu et al. presented DNN for the modelling of complex and nonlinear features
  - 92.1% accuracy → for the five-layered DNN with MFCC features [11].

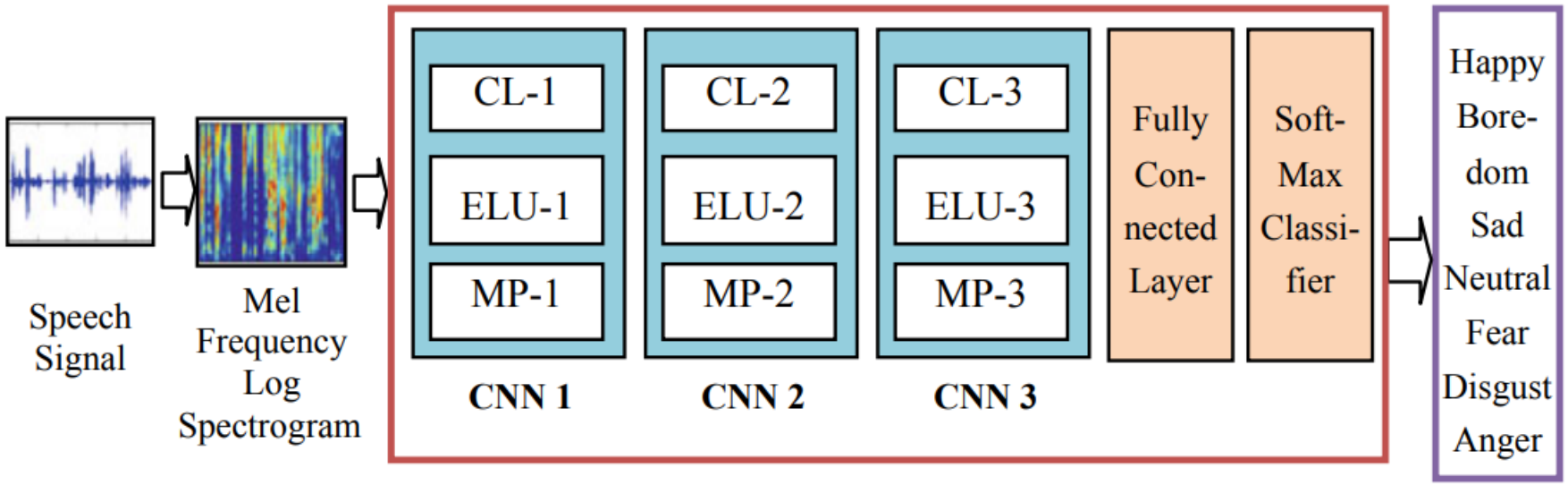
- Jianfeng Zhao et al. presented 1D CNN-long short-term memory (1D-CNN-LSTM) and 2D CNN-LSTM (2D-CNN-LSTM) to discover local and global emotion-specific features.
  - It resulted in 95.33 and 95.89% accuracy on the Berlin Emo-DB database for speaker-dependent (SD) and speaker-independent (SID) approaches [15]

- Huang et al. presented CNN for the representation of salient hierarchical features in two stages.
  - It achieved better accuracy for speaker dependent approach [12].

- Zheng et al. presented deep CNN (DCNN) for SER, which uses PCA for dimension reduction and interference suppression of the input log spectrogram.
  - It resulted in an accuracy of 40% for IEMOCAP database [13].

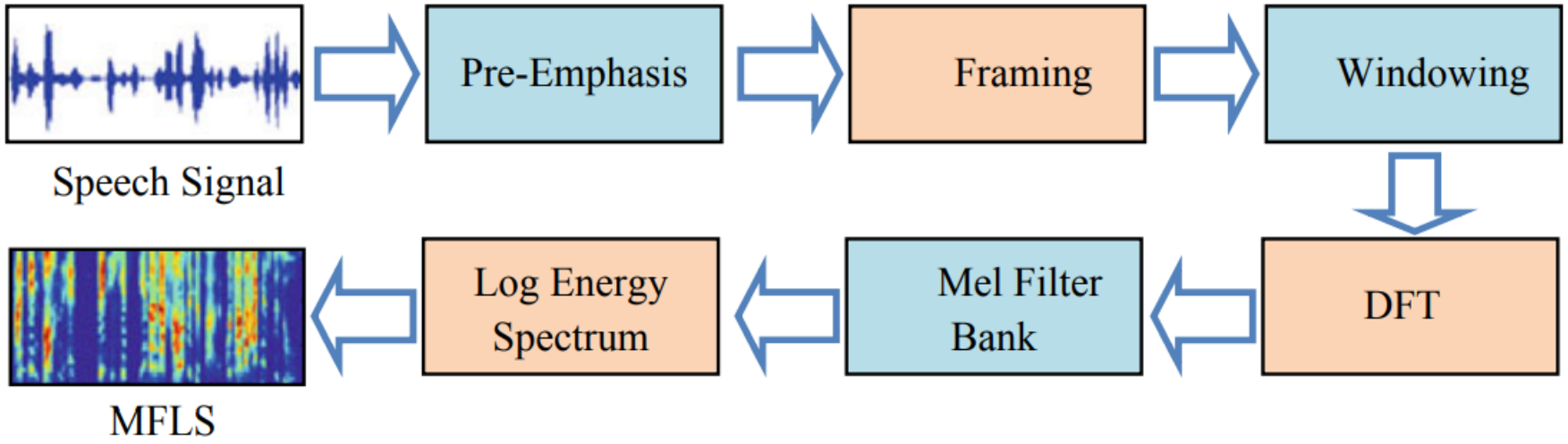
- Abdul Malik Badshah et al. presented an SER system based on DCNN with three fully connected layers that used spectrogram.
  - It resulted in 84.3% accuracy of the Berlin Emotion database [14].

Proposed Methodology



Mel Frequency Log Spectrogram (MFLS)

- input emotion speech signal are converted into two-dimensional mel frequency logarithmic spectrum.
- Mel frequency gives the relation between the human ear and sound perception frequency.



- Pre-emphasis.
  - The pre-emphasis filter suppresses the random noise and amplifies the high-frequency components of the speech emotion signal.

- Framing and Windowing.
  - The speech emotion signal is non-stationary , to process stable speech components, it is split into frames of the 40 ms.
  - Hamming window is used to collect the closest frequency components together and avoid the leakage phenomenon.

- Discrete Fourier Transform (DFT).
  - DFT is used to transform the time-domain speech emotion signal into the frequency domain.
  - Calculate emotion power spectrum P(k)

- Mel Filter Bank.
  - Calculate triangular filter bank response Hm(k)
  - The mel spectrum can be obtained by passing the emotion power spectrum P(k) through the mel-scale triangular filter bank. The product of P(k) and Hm(k) is computed at each frequency.

- Mel Frequency Logarithmic Spectrum.
  - Calculate The logarithmic energy spectrum S(m) by P(k) and Hm(k) for each frame.

Deep Convolutional Neural Network

- Convolution Layer.
  - describes the spatial local connectivity and correlation of the local region of the mel frequency spectrogram
  - two-dimensional mel frequency spectrogram S(m) convolved with the convolution kernel w(i, j).

- Exponential Linear Unit (ELU).
  - ELU removes the negative weights from the convolution layer output and normalizes the convolution layer output.

- Maximum Pooling Layer (MP).
  - acts as a nonlinear function.
  - It increases the robustness of features against distortions and noise.
  - also helps to reduce the feature dimension.

- Fully Connected Layer (FC).
  - It combines each neuron of a single layer to every neuron of other layers.

- Softmax Classifier.
  - Softmax classifier is used for the multiclass emotion classification, which is the generalized framework of the logistic regression. Softmax function provides the probability of the predicted class (Pi).

Table 3 Comparison of the proposed method with the previous implementation based on % accuracy (Emo-DB)

Research work	Method	Speaker-dependent approach	Speaker-independent approach
Huang et al. [12]	CNN	88.30	85.20
Zhao et al. [15]	CNN-LSTM	95.33	95.89
Proposed work	MFLS + DCNN	95.68	96.07

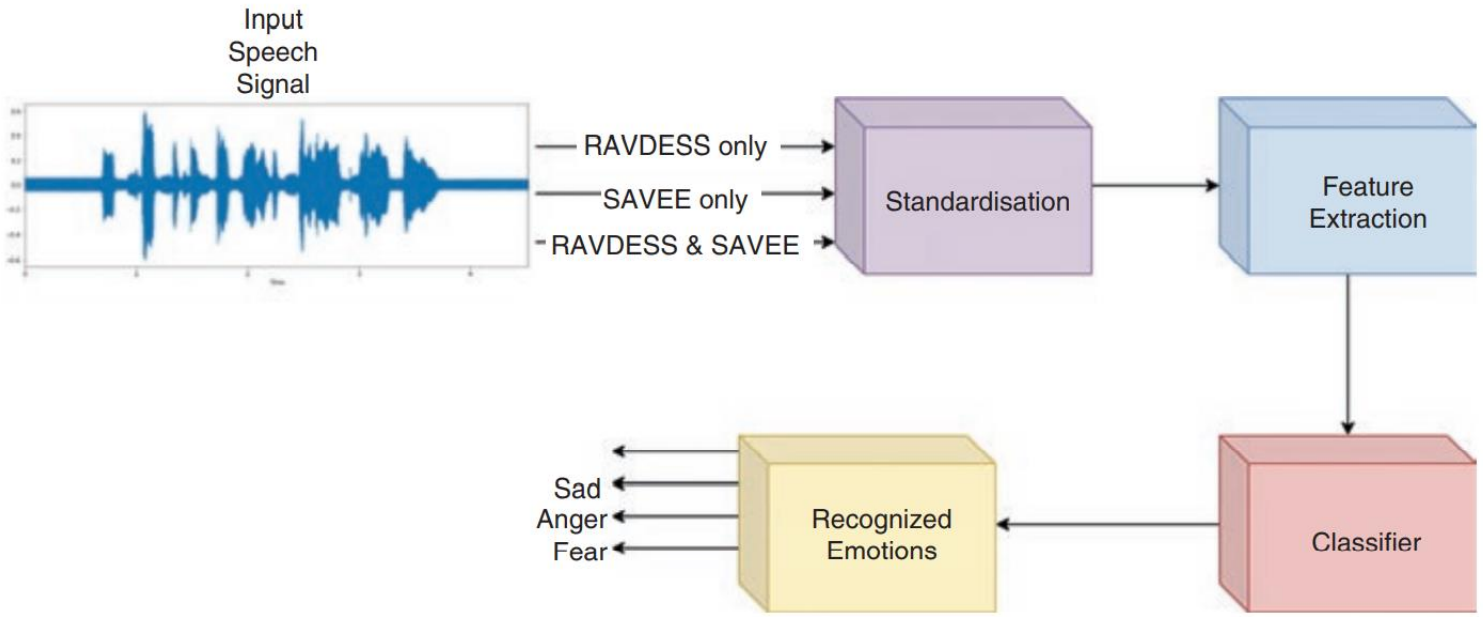
2-Emotion Recognition from Speech Signals Using Machine Learning and Deep Learning Techniques

Six models were trained and tested on both datasets individually and combined.

Dataset: RAVDESS and SAVEE

Methodology:

1-pre-processed using standardisation technique.
2-features extraction
3-classification



- Data standardisation
  - is the process of rescaling one or more attributes or features so that they'll have the properties of Gaussian distribution with a mean value of 0 and a standard deviation of 1.

- features extraction
  - spectral features → 40 MFCC, 12 chroma features, spectral centroid, etc.
  - prosodic features → zero-crossing rate, root mean square, etc.

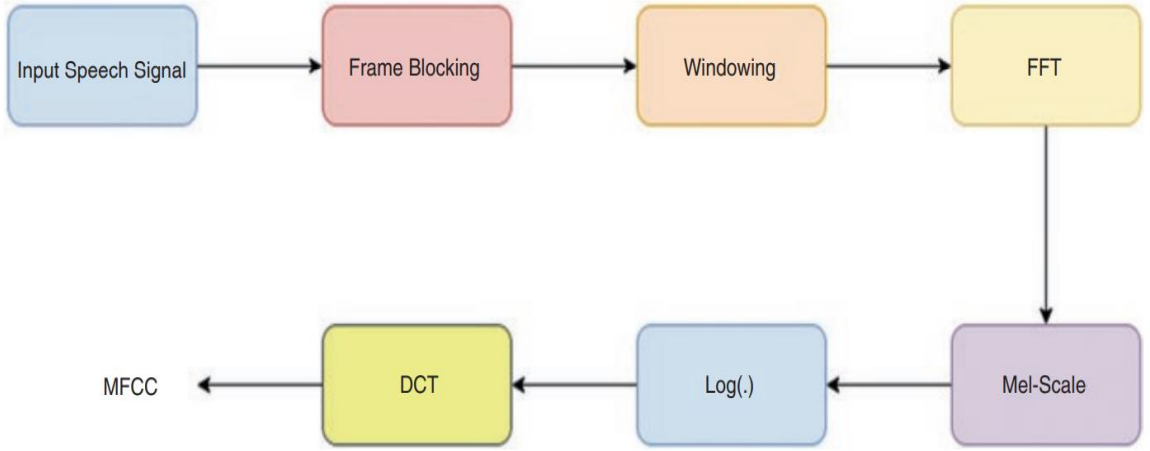


Fig. 2 The procedure to obtain MFCC features

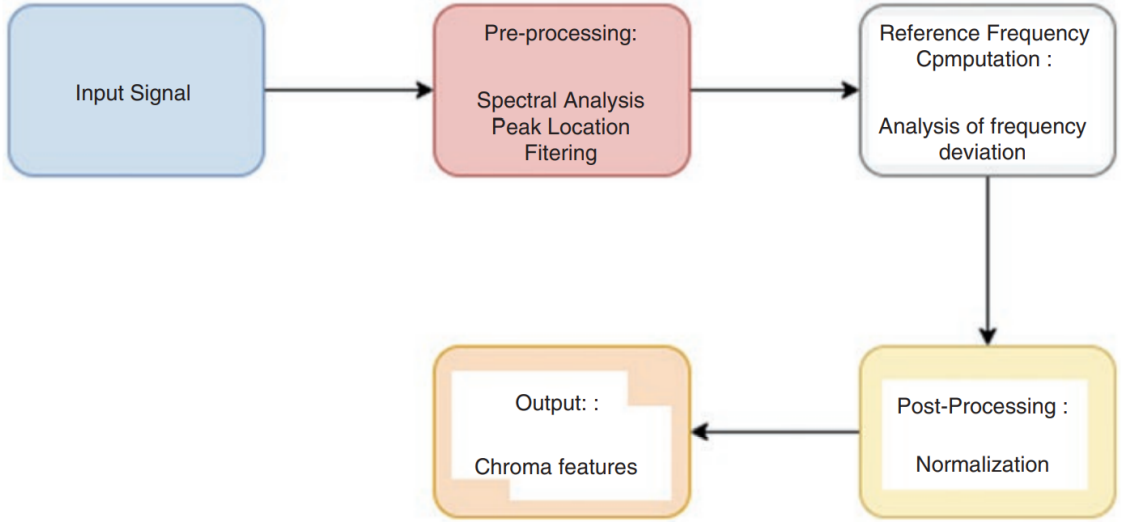


Fig. 3 The block diagram of the procedure of chroma feature extraction

Techniques:

1-support vector machine (SVM) algorithm

is most effective for classifcation problems.

It can be used for linear as well as non-linear classifcations depending on the different kernel functions.

2-Multilayer Perceptron Neural Network (MLPNN)

is a feedforward network and is used for classifcation problems.

This type of neural network has been used widely in speech recognition and image recognition

3-convolutional neural network

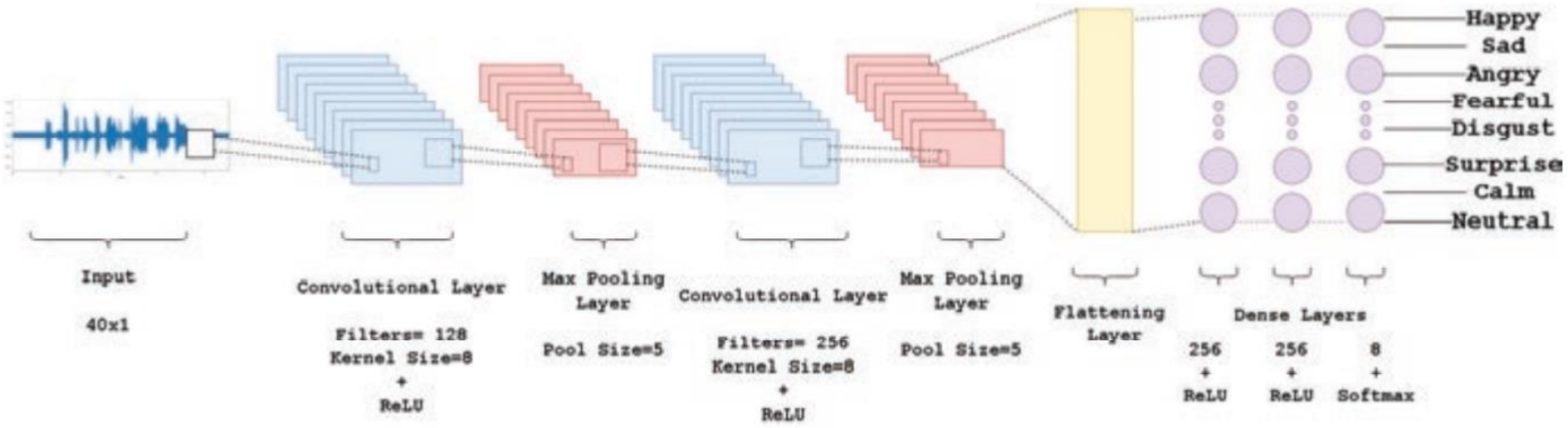


Fig. 4 The architecture of the CNN model

Classifiers	Classification accuracy (RAVDESS)
Logistic regression	52.78%
Random forest	61.81%
MLPNN	68.75%
CNN	70.83%
SVM	70.13%
KNN	61.11%

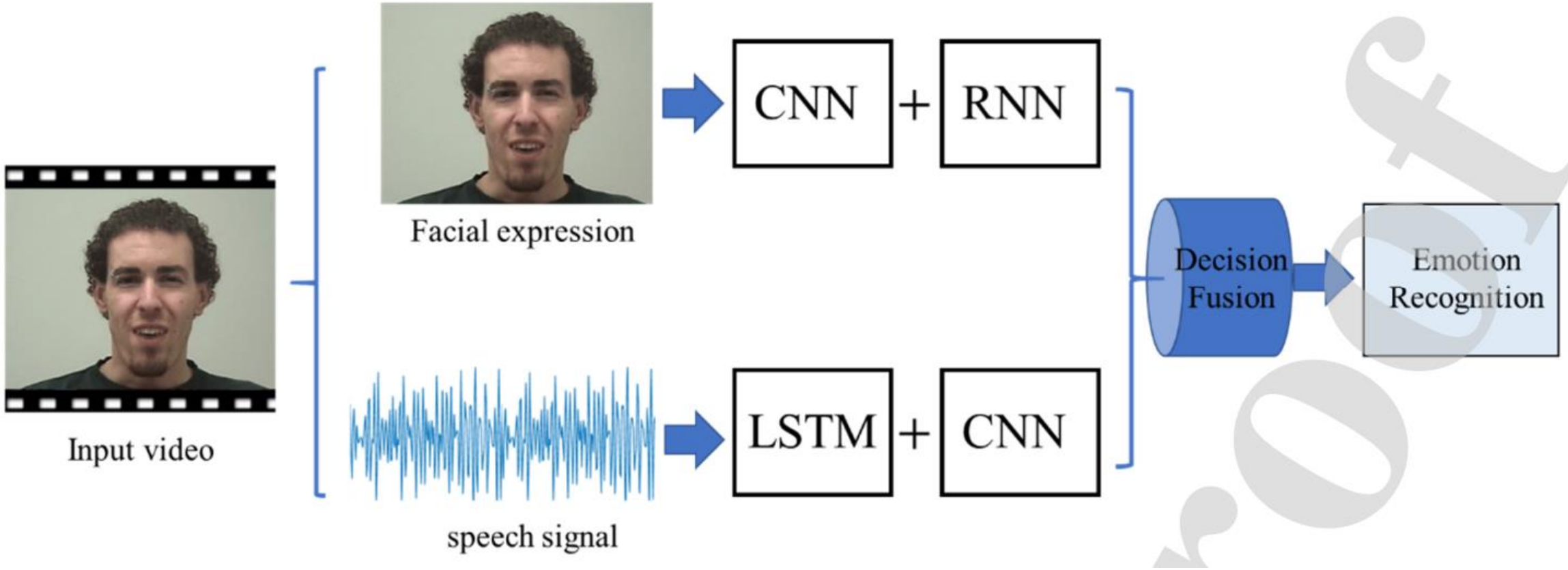
Classifiers	Classification accuracy (RAVDESS+SAVEE)
Logistic regression	44.01%
Random forest	60.68%
MLPNN	68.22%
CNN	65.46%
SVM	68.22%
KNN	58.33%

Classifiers	Classification accuracy (SAVEE)
Logistic regression	70.83%
Random forest	68.75%
MLPNN	75.00%
CNN	61.46%
SVM	70.83%
KNN	63.54%

Classifiers	Binary classification accuracy (RAVDESS)
Logistic regression	93.51%
KNN	93.5%
SVM	96.10%
MLPNN	90.55%

3-Human emotion recognition by optimally fusing facial expression and speech feature.

In our work, we combine both the facial expression and speech information to achieve emotion recognition.



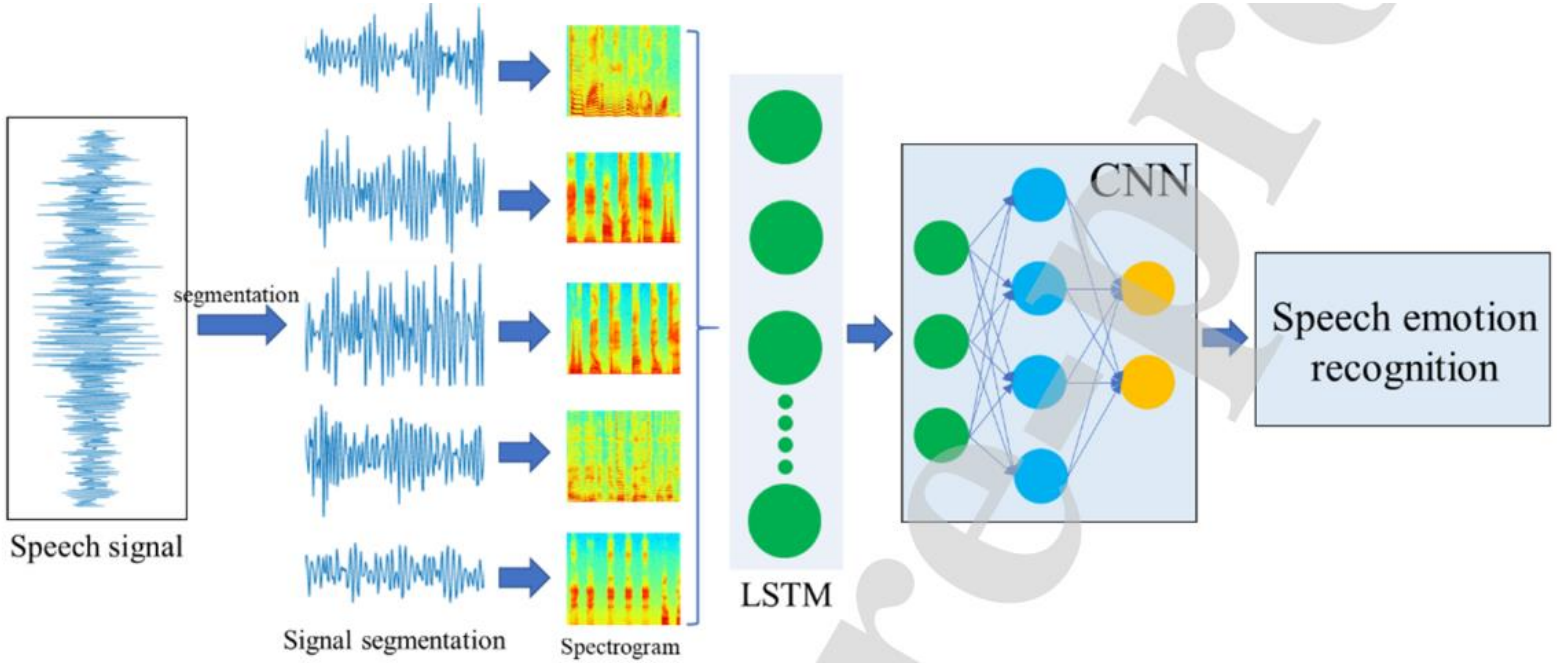
Speech emotion recognition

- ❑ In order to achieve speech emotion recognition, speech signal should be converted into images such as the Mel Frequency Cepstrum Coefficient (MFCC).
- ❑ combine the LSTM and CNN to achieve speech emotion recognition.

1. features extracted by MFCC.
2. LSTM architecture learn the temporal correlation of speech sequences.
3. The features extracted by the LSTM is fed into the CNN architecture
4. and we utilize the Softmax function to achieve speech emotion classification/retrieval.

MFCC

- divide signal into multiple equal-length segments
- each segment is transformed into frequency-domain and processed by Mel filters.



Bimodal fusion for emotion recognition

- we integrate both the facial expression and speech signal for speech emotion recognition.
- In our implementation, both the human facial expression recognition and speech emotion recognition use the Softmax function for classification.

speech emotion recognition as $S^{face} = \{S_1^{face}, S_2^{face}, S_3^{face}, \dots, S_k^{face}\}$ and $S^{speech} = \{S_1^{speech}, S_2^{speech}, S_3^{speech}, \dots, S_k^{speech}\}$, where k denotes the number of human emotion categories. Then, the weighted decision fusion is calculated as:

$$S = w_0 S^{face} + w_1 S^{speech} \quad (5)$$

where w_0 and w_1 denotes the two weights, and $w_0 + w_1 = 1$.

Table I. The comparison results in facial expression recognition

	RML	AFEW6.0	eNTERFACE'05
Baseline	0.7263	0.3879	0.5932
Improved AlexNet	0.8551	0.4382	0.7484
VGG-Face	0.8712	0.4593	0.7667
CNN+RNN	0.8896	0.4830	0.7838

Table II. The comparison results in speech emotion recognition

	RML	AFEW6.0	eNTERFACE'05
Baseline	0.7600	0.3092	0.4163
CNN	0.8362	0.3506	0.4686
BLSTM	0.8123	0.3436	0.4338
LSTM+CNN	0.8546	0.3790	0.4915

Table III. Weight settings on the three datasets

	Facial expression recognition	speech emotion recognition
RML	0.6	0.4
AFEW6.0	0.75	0.25
eNTERFACE'05	0.8	0.2