المدرسة الوطنية للعلوم التطبيقية بتطوان
ⵜⵉⵎⵔⵙⵜ ⵜⴰⵏⴰⵡⴰⵜⵜ ⵜⵉⵎⵙⵙⵓⵔⵉⵏ ⵜⵉⵜⵡⴰⵏ
Ecole Nationale des Sciences Appliquées de Tétouan

SPEECH EMOTION **RECOGNITION**

PREPARED BY :
TIDAADAR FATIMA EZZAHRAA
YOUNES OMAIMA
SEBBAR ASMAE

SUPERVISED BY :
DR.BELCAID ANASS

27 November 2024

TABLE OF **CONTENTS**

APPLICATIONS OF SER

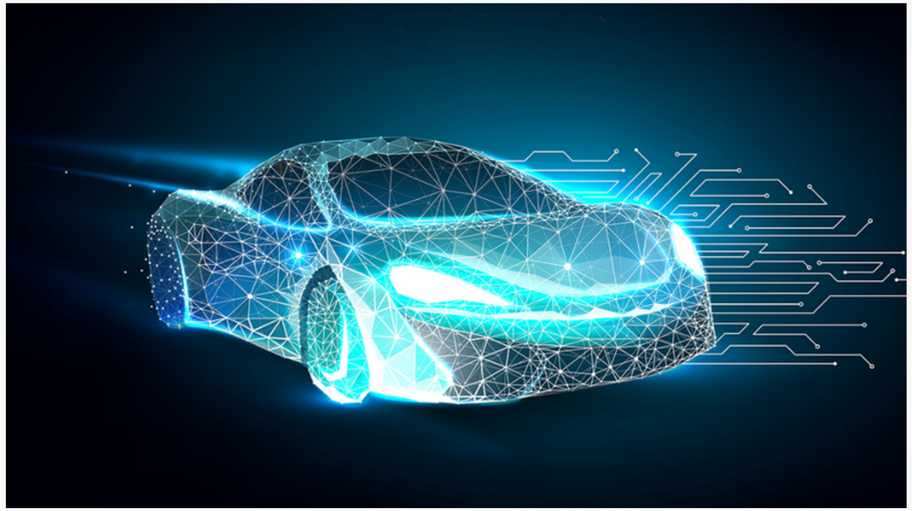
- Mental Health

- Lie Detection

APPLICATIONS OF SER

- Emergency call centers

- Automotive



DATASETS FOR SER

- **RAVDESS :** https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
- **TESS :** https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess
- **CREMA-D:** https://www.kaggle.com/datasets/ejlok1/cremad

RAVDESS

RYERSON AUDIO-VISUAL DATABASE OF EMOTIONAL SPEECHAND SONG

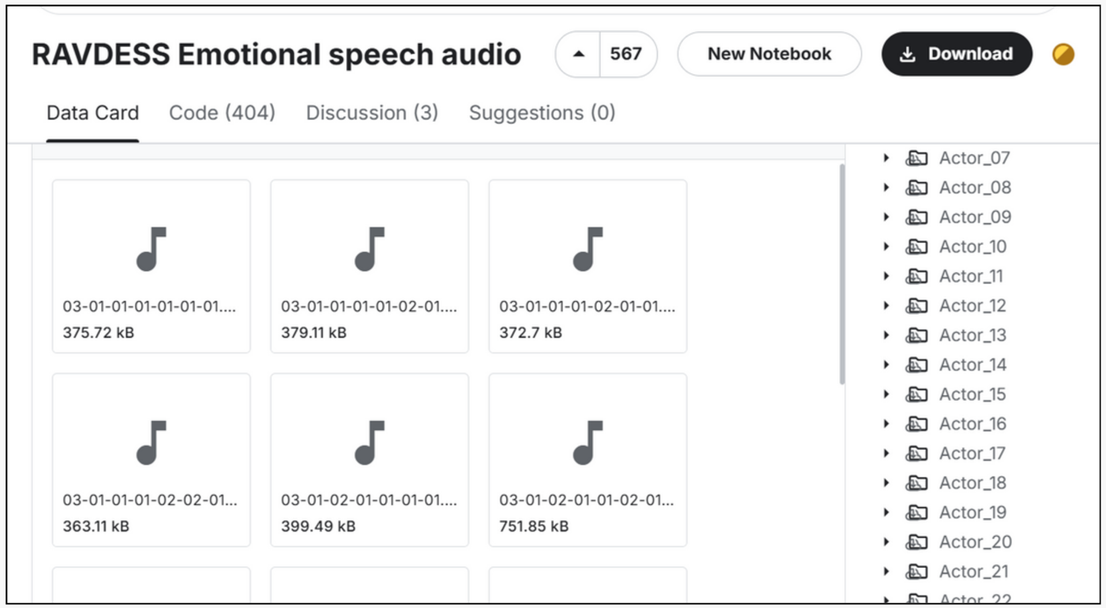
- **Type:** Audio and Video
- **Format:** WAV for audio files.
- **Quality:** 48 kHz, 16-bit for audio files.
- **Total Number of Files:** 1,440 files (720 audio + 720 video).
- **Actors:** 24 actors (12 men and 12 women).
- **Emotions:** Anger, Disgust, Fear, Joy, Sadness, Surprise, Neutral, Calm.
- **Phrases:** Two different phrases were read by the actors: "Kids are talking by the door" and "Dogs are sitting by the door".

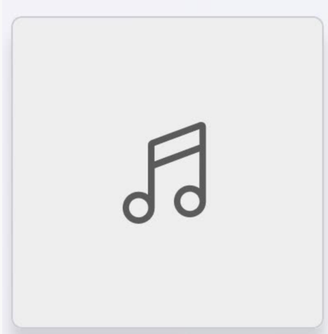
- One limitation of this dataset is that it is based solely on two sentences, which may limit the linguistic and contextual diversity of the expressed emotions.

RAVDESS

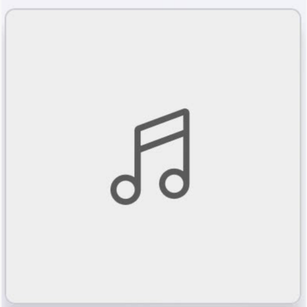
RYERSON AUDIO-VISUAL DATABASE OF EMOTIONAL SPEECHAND SONG

Example:





Happy



angry

TESS

TORONTO EMOTIONAL SPEECH SET

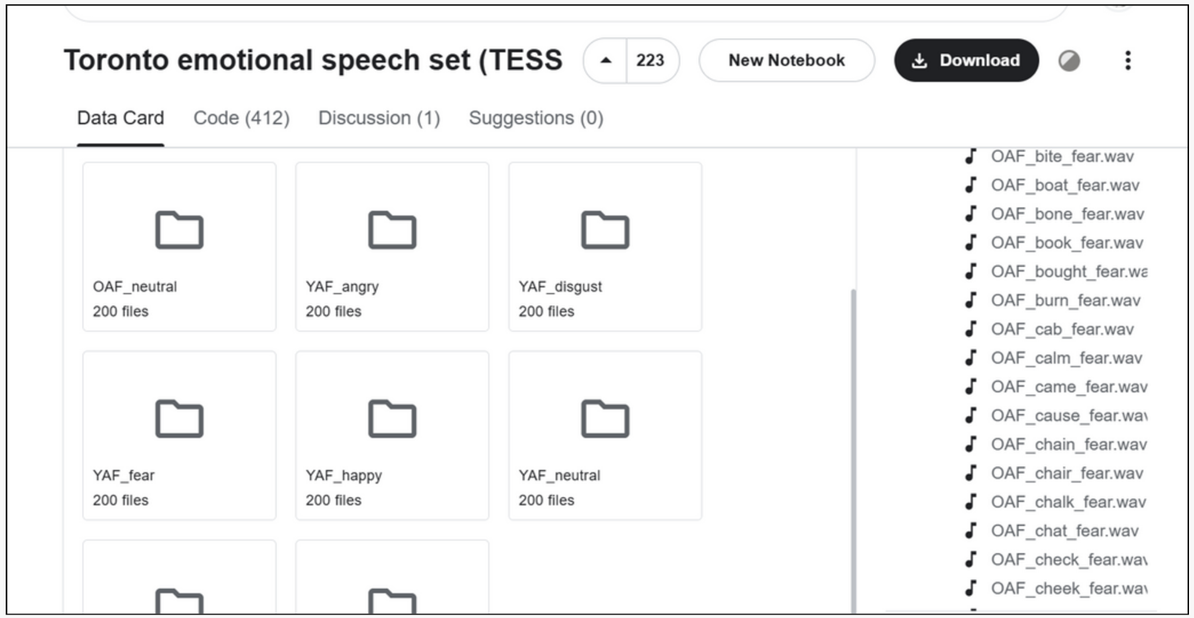
- **Type:** Audio
- **Format:** WAV
- **Quality:** 16 kHz, 16-bit for audio files
- **Total Number of Files:** 2,800 audio files
- **Actors:** 2 actresses (female)
- **Emotions:** Anger, Disgust, Fear, Joy, Sadness, Surprise, Neutral
- **Phrases:** Each actress recorded 100 phrases per emotion, for a total of 700 phrases per actress (7 emotions × 100 phrases per emotion).

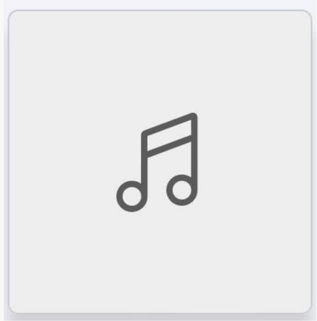
- One limitation of this dataset lies in the limited representation, as it includes exclusively female voices, which may affect the generalization of emotional analysis models.

TESS

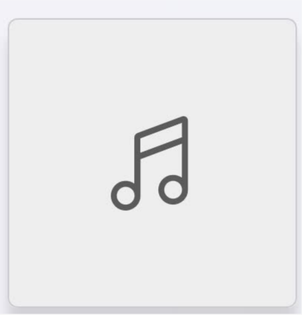
TORONTO EMOTIONAL SPEECH SET

Example:





Happy



angry

CREMA -D

CROWD-SOURCED EMOTIONAL MULTIMODAL ACTORS DATASET

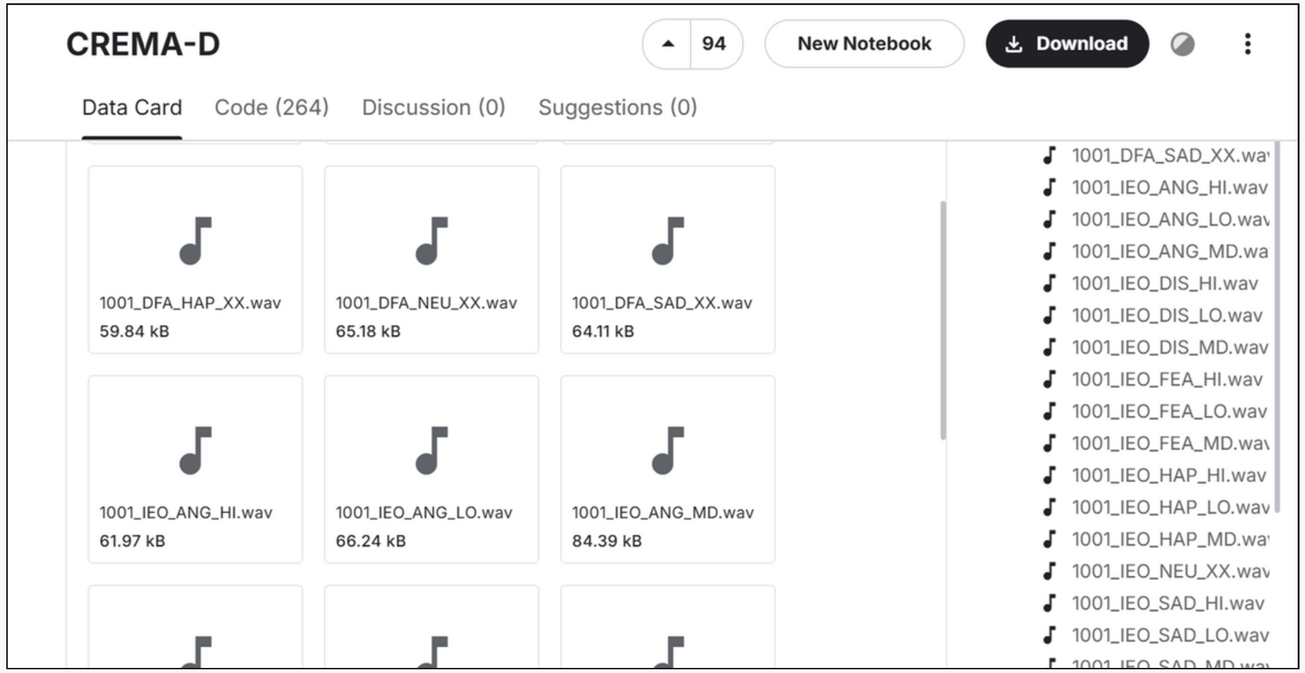
- **Type:** Audio files
- **Format:** WAV for audio
- **Quality:** 16 kHz, 16-bit for audio files
- **Total Number of Files:** **7,442** original recordings
- **Actors:** 91 actors (48 males, 43 females) aged 20 to 74, representing racial and ethnic diversity
- **Emotions:** Anger, Disgust, Fear, Joy, Neutral, Sadness
- **Phrases:** 12 distinct phrases read by actors, expressing different emotions and intensities (low, medium, high, unspecified).

- This dataset stands out for its diversity, making the trained models more robust and less likely to overfit.

CREMA -D

CROWD-SOURCED EMOTIONAL MULTIMODAL ACTORS DATASET

Example:



Happy

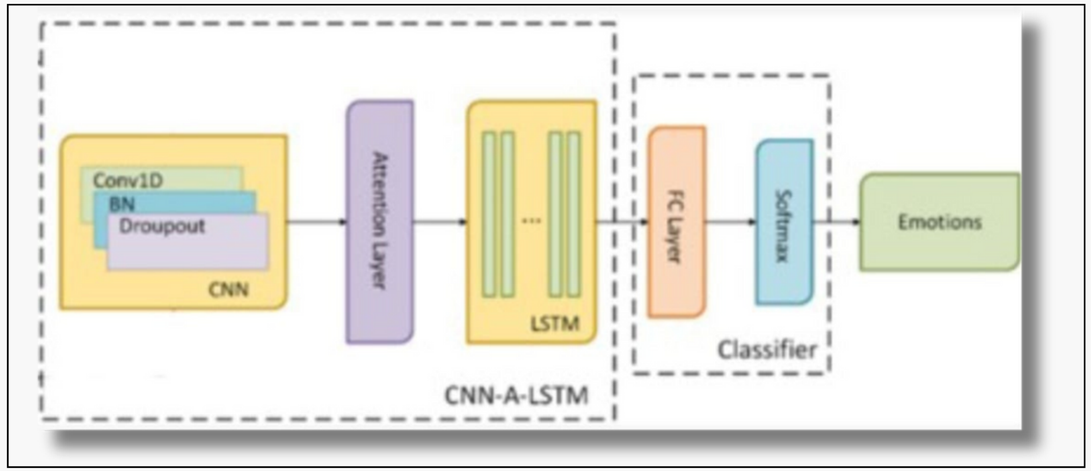
angry

ALGORITHMS FOR SER

- Algorithme n°1: CNN+LSTM
- Algorithme n°2: CNN + BiLSTM
- Algorithme n°3: Wav2Vec2 with Sequence Classification
- Algorithme n°4: LFPC + HMM
- Algorithme n°5: HMM + SVM

CNN & LSTM

- **Conv1D:** Captures local patterns in audio sequences.
- **Batch Normalization (BN):** Stabilizes and accelerates training.
- **Dropout:** A regularization technique used to prevent overfitting by randomly deactivating a percentage of the neurons during training.
- **Attention Layer:** Identifies and weights important parts of the audio signal, additionally emphasizes the areas of the signal that are most significant for classification.
- **LSTM:** A Long Short-Term Memory (LSTM) recurrent neural network layer that captures the temporal dependencies in the audio data.
- **Fully Connected Layer:** Maps the representations extracted by the CNN and LSTM to an emotion space. It uses non-linear activations (e.g. ReLU).
- **Softmax Classifier:** Provides a final classification in probabilities for each emotion (e.g. joy, sadness, anger).



https://link.springer.com/article/10.1007/s11042-023-17829-x

CNN & BILSTM

1)-Input(Extracted Speech) : The raw speech signal is processed through two paths: **Mel-Spectrogram (left)** and **Feature Set (right).**

2)-Left Path: Mel-Spectrogram and CNN-based Processing:

- **Mel-Spectrogram Conversion:**Converts the speech signal into a time-frequency representation.
- **Convolutional Neural Network (CNN) Layers:** Applies four convolutional layers with filters of sizes: 32, 64, 128, and 256 (all 3×3).
- **Pooling Layers:**Combines Global Average and Max Pooling results into a feature vector **(Output 1)**.

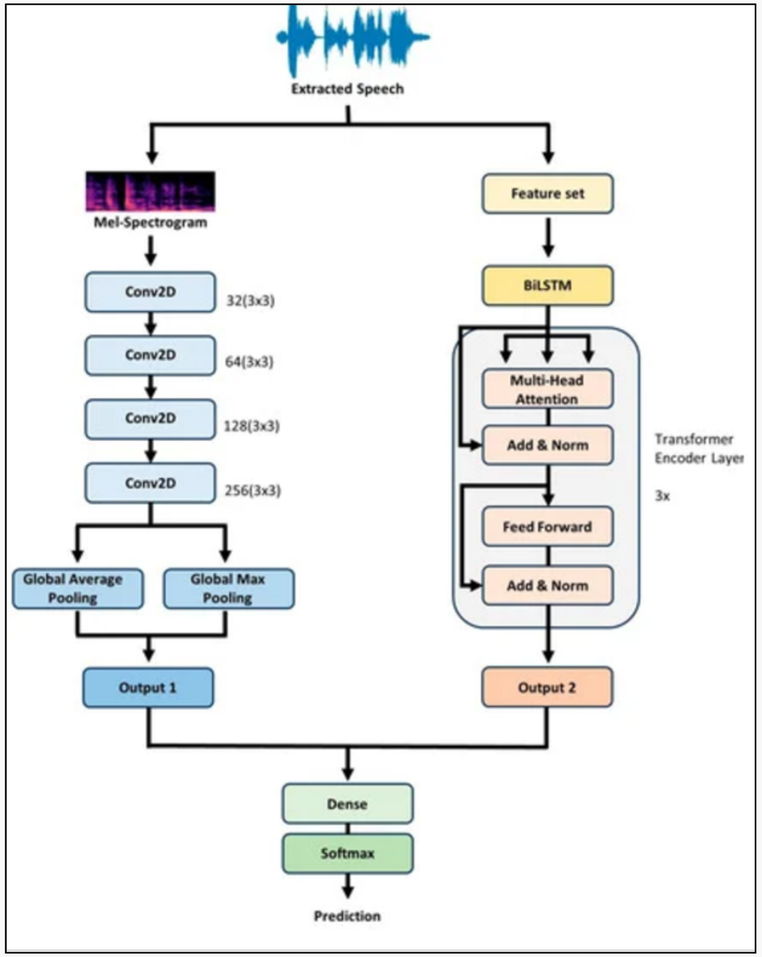


Figure 3. BiLSTM–Transformer + 2D CNN model architecture.
Paper:https://www.mdpi.com/2079-9292/12/19/4034

CNN & BILSTM

3)-Right Path: Feature Set and BiLSTM with Transformer Layers:

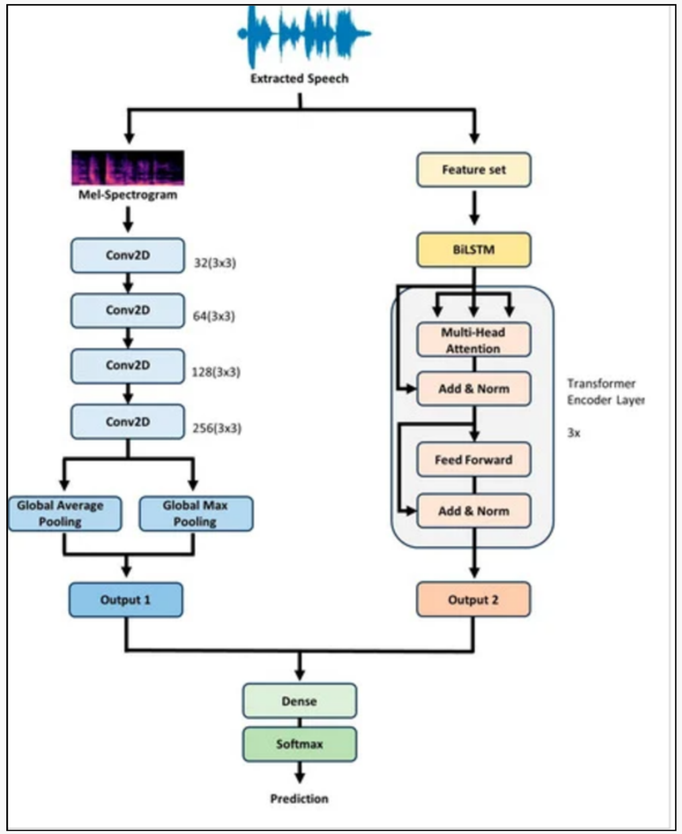
Feature Set: Extracted audio features (MFCCs, spectral features) summarize the speech signal.

Bidirectional LSTM (BiLSTM): Captures temporal patterns by analyzing the sequence in both forward and backward directions.

Transformer Encoder Layers:

- **Multi-Head Attention:** Highlights key parts of the input.
- **Add & Norm Layers:** Stabilizes inputs.
- **Feed-Forward Network:** Processes features further.
- Repeated 3 times for enhanced representation.

Output 2: Outputs a feature vector.



4)-Merging Paths:

Output 1 (CNN) + Output 2 (BiLSTM + Transformer) are combined into a unified feature representation.

5)-Dense Layer and Softmax for Classification

The combined feature representation is passed through:

- **Dense Layer:** Refines combined features.
- **Softmax Layer**: Predicts emotion probabilities as the final output

WAV2VEC2 WITH SEQUENCE CLASSIFICATION

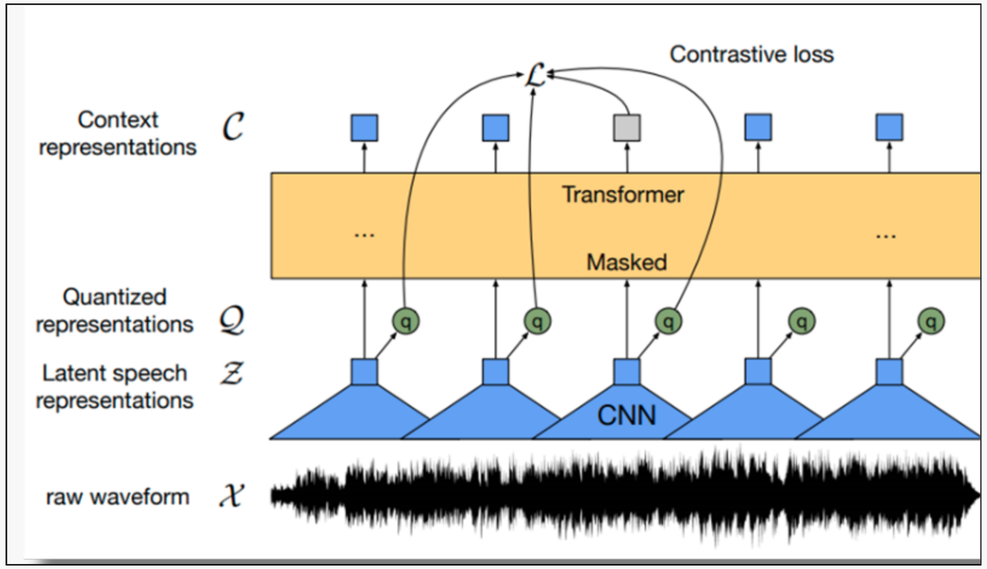
Architecture : CNN + Transformer + Sequence classification Head

-Wav2Vec2 (Modèle pré-entrainé par Facebook AI):

- Feature Extraction with CNN: Uses convolutional networks (CNNs) to transform the raw sound wave into compact latent representations, which capture local acoustic properties.
- Quantization: Latent representations are quantized to reduce dimensionality.
- Transformer for Context Representations (C): A part of the quantized representations is hidden, and the Transformer is responsible for predicting the missing values.

-Classification head (SequenceClassification):

- A fully connected (dense) layer is added to map high-level representations to specific classes (like emotions: happiness, sadness, fear, etc.). It can include a Softmax layer to obtain class probabilities.



https://link.springer.com/article/10.1007/s11042-023-17829-x

ALGORITHMS AND ACCURACY FOR SER

Algorithm	Accuracy
CNN + LSTM	**82.4%**
CNN + BiLSTM	**82.31%**
Wav2vec2 with Sequence Classification	**99.6%**
LFPC + HMM	**78%**
HMM + SVM	**82.14 %**

CHALLENGES

- **Linguistic diversity:** Pronunciation, languages, and accents impact model generalization.
- **Cultural variation:** Emotions are interpreted differently across cultures.
- **Background noise:** Real-world recordings often include interferences.
- **Complex emotions:** Emotions like jealousy are harder to identify due to varied reactions.
- **Lack of visual cues:** Absence of facial expressions limits accuracy for complex emotions.

FUTURE DIRECTIONS

- **Recognition of Subtle Emotions:** detect more subtle emotions, such as emotional ambiguity, anxiety, or even mixed emotions where multiple feelings are expressed simultaneously.
- **Contextual and Cultural Adaptation:** adapt to specific cultural contexts and understand the nuances of emotions across various languages and dialects.
- **Real-Time Detection for Live Interactions:** Real-time emotion detection systems are increasingly being integrated into live interactions, such as phone calls or video conferences.
- **Applications in Mental Health:** helps to identify early signs of conditions such as depression, anxiety, or stress.

THANK YOU !