Emotion Recognition in Speech Using Deep Learning

**Abhinav Choudhary [1] B. Tech CSE *Manbir Kaur [2] 293393

12015798 RKE030B41 Assistant Professor

Lovely Professional University, Phagwara Punjab.

Abstract: Emotion recognition in speech through deep learning methods has become increasingly significant within the domain of Human-Computer Interaction. This research delves into the application of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), utilizing renowned libraries like TensorFlow and Keras, alongside the librosa library for feature extraction. The spectrograms generated via speech stretching serve as input features for the models. Remarkably, the study achieves an accuracy rate exceeding 90% in emotion recognition tasks. These advancements carry substantial implications for diverse applications, encompassing affective computing, virtual assistants, and mental health monitoring. Moreover, the paper outlines future directions for refining emotion recognition systems, including exploring multimodal approaches and addressing real-time processing challenges. This study underscores the crucial role of deep learning in enabling precise and effective emotion recognition from speech, thereby fostering more sophisticated human-computer interactions and innovative technological solutions.

Keywords: Emotion recognition, Speech analysis, Deep learning, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Librosa, TensorFlow, Keras, Spectrogram, Feature extraction, Accuracy, Human-Computer Interaction (HCI), Affective computing, Multimodal approaches, Real-time processing, Technological solutions.

Introduction: Emotion recognition in speech has become a crucial part of Human-Computer Interaction (HCI) [1-3]. The purpose of these systems is to enable the natural voice-to-machine interface that replaces the need for conventional devices as input to comprehend spoken material and facilitate human listeners' responses [4-6]. Applications for spoken language dialogue systems include call center conversations, onboard car navigation systems, and the use of spoken emotion patterns in medical applications [7]. However, there are still a lot of issues with HCI systems that need to be adequately resolved, especially as these systems transition from laboratory testing to practical use [8]– [10]. Therefore, work is needed to develop better emotion identification robots and find effective solutions to such issues.

Human emotional state determination is a unique problem that can serve as a benchmark for any algorithm that recognizes emotions [11]. One of the many models that are employed to classify different emotions is a discrete emotion oversaw the manuscript's evaluation and gave it the go-ahead to publish. Approach is regarded to be one of the core strategies. It makes use of a range of feelings, including neutral, sadness, joy, happiness, surprise, rage, boredom, disgust, and fear [12], [13]. A three-dimensional continuous space with properties like arousal, valence, and potency is another prominent model that is used.

The feature extraction and features classification phases make up the two main stages of the speech emotion recognition (SER) technique [14]. Researchers have developed a few features in the field of speech processing, including vocal traction factors, prosodic features, source-based excitation features, and other hybrid features [15]. Using both linear and nonlinear classifiers, feature classification is done in the second phase. Support vector machines (SVM) and Bayesian networks (BN), often known as the Maximum Likelihood Principle (MLP), are the most widely used linear classifiers for emotion recognition. The speech signal is typically regarded as non-stationary. Non-linear classifiers are therefore thought to function well for SER. For SER, a variety of non-linear classifiers are available, including the Hidden Markov Model (HMM). These are widely used for classification of information that is derived from basic level features.

Deep Learning has emerged as a prominent research area within machine learning, garnering increased attention in recent years. Its application in Speech Emotion Recognition (SER) offers distinct advantages over traditional methods, notably its ability to automatically detect complex structures and features without the need for manual extraction and tuning. Additionally, Deep Learning techniques excel in extracting low-level features from raw data and handling unlabeled datasets.

Deep Neural Networks (DNNs) form the backbone of many Deep Learning architectures, featuring feed-forward structures with hidden layers between inputs and outputs. Architectures like DNNs and Convolutional Neural Networks (CNNs) have demonstrated efficiency in tasks such as image and video processing. Conversely, recurrent architectures like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are particularly effective in speech-based classification tasks like natural language processing (NLP) and SER.

However, these models also exhibit limitations. While CNNs excel in learning features from high-dimensional data, they may inadvertently capture small variations and distortions, necessitating substantial storage capacity. Similarly, LSTM-based RNNs are proficient in handling variable input data and modeling long-range sequential text data, yet they too have their constraints.

TRADITIONAL TECHNIQUES FOR SER

Emotion recognition systems built upon digitized speech encompass three fundamental components: signal preprocessing, feature extraction, and classification. Acoustic preprocessing, including denoising and segmentation, is conducted to identify meaningful segments within the signal. Feature extraction aims to discern pertinent features present in the signal, while classifiers map extracted feature vectors to relevant emotions. This section provides an in-depth exploration of speech signal processing, feature extraction, and classification. Additionally, it discusses the distinctions between spontaneous and acted speech, considering their relevance to the subject. Figure 1 illustrates a simplified system employed for speech-based emotion recognition. In the initial stage of speech signal processing, noise reduction is executed to eliminate noisy elements.

Fig 1: Traditional Speech Recognition System

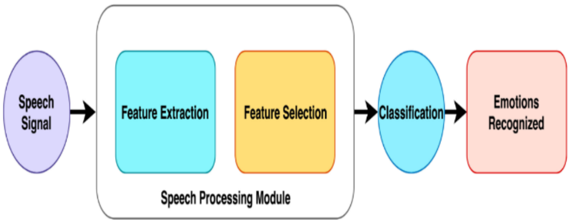
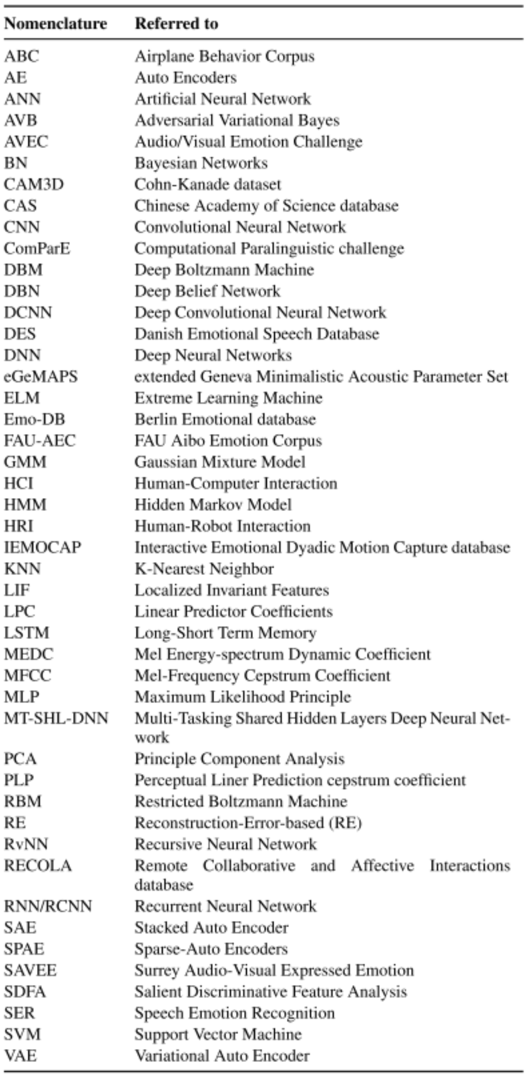


Table1: List of nomenclature in this paper



Nomenclature	Referred to
ABC	Airplane Behavior Corpus
AE	Auto Encoders
ANN	Artificial Neural Network
AVB	Adversarial Variational Bayes
AVEC	Audio/Visual Emotion Challenge
BN	Bayesian Networks
CAM3D	Cohn-Kanade dataset
CAS	Chinese Academy of Science database
CNN	Convolutional Neural Network
ComParE	Computational Paralinguistic challenge
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DCNN	Deep Convolutional Neural Network
DES	Danish Emotional Speech Database
DNN	Deep Neural Networks
eGeMAPS	extended Geneva Minimalistic Acoustic Parameter Set
ELM	Extreme Learning Machine
Emo-DB	Berlin Emotional database
FAU-AEC	FAU Aibo Emotion Corpus
GMM	Gaussian Mixture Model
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
HRI	Human-Robot Interaction
IEMOCAP	Interactive Emotional Dyadic Motion Capture database
KNN	K-Nearest Neighbor
LIF	Localized Invariant Features
LPC	Linear Predictor Coefficients
LSTM	Long-Short Term Memory
MEDC	Mel Energy-spectrum Dynamic Coefficient
MFCC	Mel-Frequency Cepstrum Coefficient
MLP	Maximum Likelihood Principle
MT-SHL-DNN	Multi-Tasking Shared Hidden Layers Deep Neural Network
PCA	Principle Component Analysis
PLP	Perceptual Liner Prediction cepstrum coefficient
RBM	Restricted Boltzmann Machine
RE	Reconstruction-Error-based (RE)
RvNN	Recursive Neural Network
RECOLA	Remote Collaborative and Affective Interactions database
RNN/RCNN	Recurrent Neural Network
SAE	Stacked Auto Encoder
SPAE	Sparse-Auto Encoders
SAVEE	Surrey Audio-Visual Expressed Emotion
SDFA	Salient Discriminative Feature Analysis
SER	Speech Emotion Recognition
SVM	Support Vector Machine
VAE	Variational Auto Encoder

The subsequent stage comprises feature extraction and selection, involving the extraction of requisite features from the preprocessed speech signal and subsequent selection based on analysis in the time and frequency domains. Finally, various classifiers such as GMM and HMM are employed for feature classification, ultimately facilitating the recognition of different emotions.

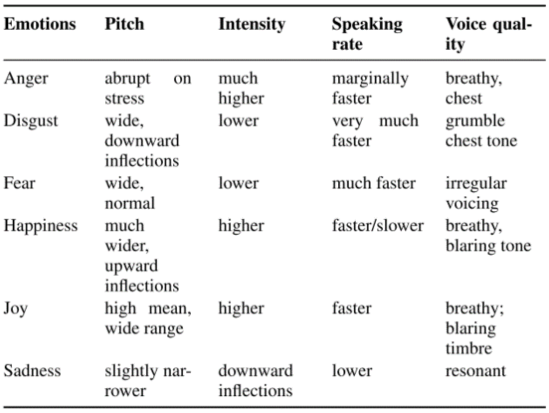
1.ENHANCEMENT OF INPUT SPEECH DATA

Improving the quality of input speech data in Speech Emotion Recognition (SER) systems is crucial for accurate analysis. Noise contamination during data capture can significantly impair feature extraction and classification accuracy. Therefore, enhancing the input data is essential for effective emotion detection and recognition. During the preprocessing stage, efforts are made to preserve emotional distinctions while mitigating variations introduced by different speakers and recording conditions. This ensures that the input data maintains its integrity and fidelity for subsequent analysis.

2. FEATURE EXTRACTION AND SELECTION

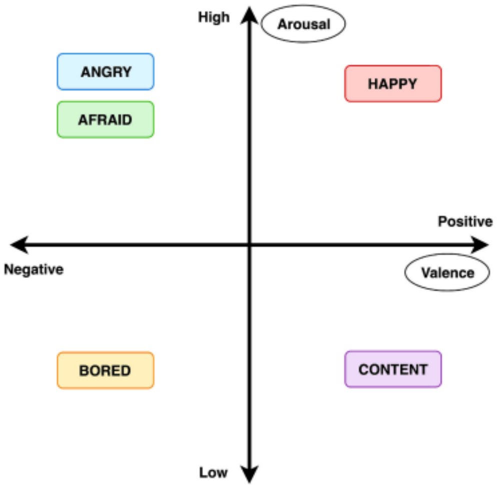
Feature extraction and selection play pivotal roles in Speech Emotion Recognition (SER). Following the enhancement of the speech signal, it is segmented into meaningful units. Relevant features are then extracted and categorized based on the information obtained. Short-term classification relies on characteristics such as energy, formants, and pitch over brief periods. In contrast, long-term classification utilizes features such as mean and standard deviation. Prosodic features, including intensity, pitch, speech rate, and variance, are particularly significant for discerning different emotions within the input speech signal. Feature extraction is a crucial step in many signal processing and pattern recognition tasks, including Speech Emotion Recognition (SER). It involves transforming raw data, such as speech signals, into a representative set of features that capture relevant information for subsequent analysis.

Table 2: Different emotions



Emotions	Pitch	Intensity	Speaking rate	Voice quality
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

Fig 2: A 2D basic emotional space

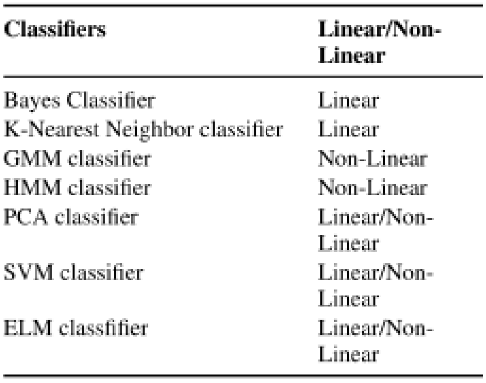


3. Classification of features

A variety of classifiers have been explored in the development of systems like Speech Emotion Recognition (SER), speech recognition, and speaker verification. However, the rationale behind selecting a particular classifier for a specific speech task is often overlooked in many applications. Typically, classifiers are chosen based on either heuristics or empirical evaluation of certain indicators.

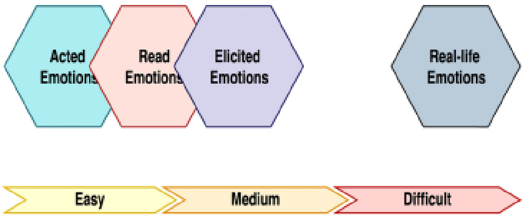
Pattern recognition classifiers used for SER can generally be categorized into two main types: linear classifiers and non-linear classifiers. Linear classifiers typically perform classification based on the arrangement of object features in a linear fashion. These features are typically evaluated in the form of a feature vector. On the other hand, non-linear classifiers are employed for object characterization by establishing non-linear weighted combinations of such features.

Table 3: Linear and non-Linear Classifiers for SER



Classifiers	Linear/Non-Linear
Bayes Classifier	Linear
K-Nearest Neighbor classifier	Linear
GMM classifier	Non-Linear
HMM classifier	Non-Linear
PCA classifier	Linear/Non-Linear
SVM classifier	Linear/Non-Linear
ELM classfifier	Linear/Non-Linear

Fig 3: Emotion Recognition databases difficulty level



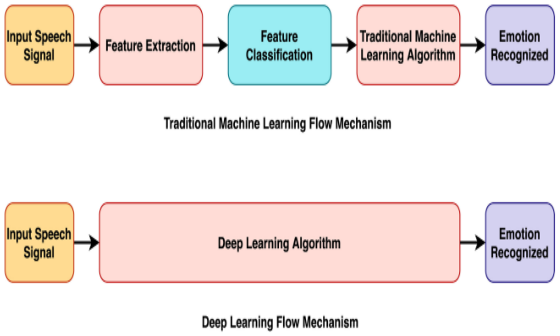
NEED OF DEEP LEARNING TECHNIQUES

Speech processing is integral to various applications such as Speech Emotion Recognition (SER), speech denoising, and music classification. While SER has gained significance, achieving human-like interaction still demands accurate methodologies. A typical SER system comprises feature selection and extraction, feature classification, acoustic modelling, recognition per unit, and language-based modelling. Traditional SER systems often integrate classification models like Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). GMMs illustrate acoustic features, while HMMs handle temporal variations in speech signals.

Deep learning methods feature nonlinear components operating in parallel . However, to surpass limitations of other techniques, deeper architectures are necessary. Techniques such as Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RNN), Deep Belief Network (DBN), Convolutional Neural Networks (CNN), and Auto Encoder (AE) are fundamental deep learning methods for SER, significantly enhancing system performance.

Deep learning is a burgeoning field in machine learning, attracting considerable attention in recent years. Some researchers employ Deep Neural Networks (DNNs) to train SER models. Particularly the Deep Convolutional Neural Network (DCNN), in measuring various emotions using datasets like IEMOCAP, Emo-DB, and SAVEE. Emotions such as happiness, anger, and sadness are recognized, with deep learning algorithms demonstrating superior performance compared to traditional techniques. The subsequent section aims to delve into various deep learning techniques in the context of SER, offering accurate results albeit with computational complexity.

Fig 4: Traditional Machine Learning Flow vs Deep Learning Flow



Convolutional Neural Network (CNN)

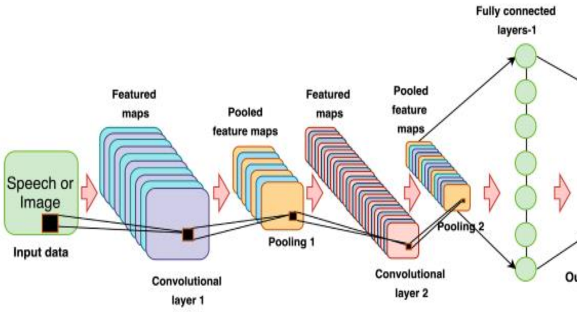
Convolutional Neural Networks (CNNs) represent another variant of Deep Learning techniques, relying on a feed-forward architecture for classification [74]. Widely employed in pattern recognition, CNNs offer superior data classification capabilities. Each layer in CNNs comprises small-sized neurons, processing input data in the form of receptive fields. Figure 5 illustrates the layer-wise architecture of a basic CNN network.

Filters serve as the foundation of local connections, convolved with the input and sharing the same parameters (weight W^i bias n^i) to generate feature maps z^i, each of size a - b - 1. The convolutional layers compute the dot product between the weights and provided inputs, formulated as

$z^i = g(W^i * r + n^i)$

Activation functions ensure non-linear output from the convolution layers. Inputs are represented by small regions of the original volumes, and down-sampling is conducted at each subsampling layer to reduce parameters and control overfitting, thereby enhancing the training process. Pooling is performed over p x p elements (filter size) across all feature maps. In the final stage, layers are fully connected akin to other neural networks, utilizing low-level and mid-level features to generate high-level abstractions from input speech data. The last layer, often SVM or SoftMax, computes classification scores in probabilistic terms, associating with specific classes.

Fig 5: Layer-wise convolutional neural network architecture.

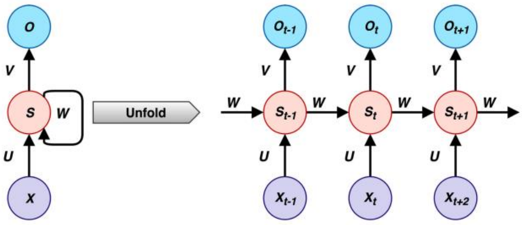


Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) represent a branch of neural networks specializing in sequential information processing, where outputs and inputs are interdependent [65]. This interdependency is particularly useful for predicting future states based on input sequences. Similar to Convolutional Neural Networks (CNNs), RNNs require memory to retain information throughout the sequential learning process and tend to work efficiently for a limited number of back-propagation steps. In the context of speech emotion recognition, RNNs are well-suited due to their ability to frame acoustic features at short-time intervals.

However, a significant challenge affecting RNN performance is its sensitivity to vanishing gradients . This sensitivity arises when gradients exponentially decay during training, resulting in the loss of input information from initial stages. To mitigate this issue, Long Short-Term Memory (LSTM) networks are employed, introducing memory blocks with gated units to regulate the influx of new information. These memory blocks maintain temporal states of the network, while residual connections aid in mitigating gradient problems, particularly in deep architectures.

Fig 6: Basic architecture of Recurrent Neural Network

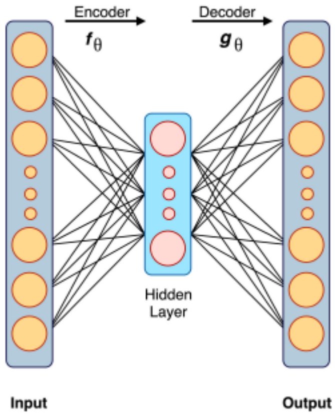


where x_t is the input, s_t is the underlaying hidden state, and o_t is the output at time step t. The U, V, W are known as parameters for hidden matrices and their values may varies for every time step. The hidden state is calculated as $St = f\,(U(x_t) + W_{s(t-1)}\,)$.

Auto Encoder

Autoencoders represent a fundamental component of Deep Learning, serving as unsupervised learning models aimed at learning efficient representations of input data. They consist of an encoder network, tasked with compressing input data into a latent space, and a decoder network, responsible for reconstructing the input from the latent space representation. By minimizing the reconstruction error between the input and the output, autoencoders learn to capture the most salient features of the data. This makes them invaluable for various tasks such as dimensionality reduction, feature learning, and data denoising. Furthermore, autoencoders have found wide-ranging applications in fields like computer vision, natural language processing, and anomaly detection. Their versatility and effectiveness make them a cornerstone of Deep Learning research and application.

Fig 9: Auto encoder architecture



Methodology

1. Data Collection and Preprocessing:

- Data Collection: Acquire a diverse dataset containing speech samples annotated with corresponding emotional labels.

- Preprocessing: Clean the dataset by removing noise, normalizing audio levels, and segmenting speech samples into manageable units.

2. Feature Extraction:

- Acoustic Features: Extract relevant acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, intensity, and spectral features from preprocessed speech signals.

- Prosodic Features: Capture suprasegmental features like pitch contour, speech rate, and intonation patterns to encode emotional prosody.

- Lexical Features: Incorporate linguistic features such as word embeddings or phonetic representations to capture semantic context.

3. Model Selection and Development:

- Deep Learning Architecture: Design a deep learning model architecture suitable for speech emotion recognition.

Consider architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or hybrid architectures like Convolutional Recurrent Neural Networks (CRNNs).

- Training Strategy: Utilize appropriate loss functions and optimization algorithms for training the model. Experiment with techniques like transfer learning or data augmentation to enhance model generalization.

- Hyperparameter Tuning: Conduct systematic experimentation to optimize model hyperparameters such as learning rate, batch size, and network architecture.

4. Model Evaluation:

- Cross-Validation: Employ k-fold cross-validation to assess model performance across different subsets of the dataset.

- Metrics: Evaluate the model using standard evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.

- Generalization: Assess the model's ability to generalize to unseen data by evaluating performance on a separate test set.

5. Comparison with Baselines:

- Baseline Models: Compare the performance of the deep learning model against baseline approaches such as traditional machine learning classifiers or rule-based systems.

- Statistical Analysis: Conduct statistical tests to determine the significance of performance differences between the proposed model and baseline methods.

6. Interpretation and Visualization:

- Model Interpretability: Explore techniques for interpreting the model's decisions, such as attention mechanisms or saliency maps, to gain insights into the learned representations.

- Visualization: Visualize learned features or embeddings using techniques like t-SNE to understand the clustering of emotional states in the latent space.

7. Ethical Considerations:

- Bias and Fairness: Assess and mitigate potential biases in the dataset or model predictions, particularly concerning sensitive attributes such as gender or ethnicity.

- Privacy: Ensure compliance with data privacy regulations and obtain informed consent from participants for data collection and usage.

Literature review

[1] B. W. Schuller, "Speech emotion recognition: Schuller provides a comprehensive overview of two decades of research in speech emotion recognition. The paper highlights benchmarks and ongoing trends in the field, shedding light on the evolution and advancements made in this domain.

[2] Hossain and Muhammad (2019)**:** Hossain and Muhammad focus on emotion recognition using deep learning approaches, particularly emphasizing the integration of audio-visual emotional big data. The study explores the potential of deep learning models in capturing complex emotional cues from multimodal data sources.

[3] Chen, Zhou, and Fortino (2016)**:** Chen, Zhou, and Fortino propose an emotion communication system aimed at enhancing emotional interaction through technological means. The paper discusses the design and implementation of a system facilitating effective communication of emotions in various contexts.

[4]Lane and Georgiev (2015)**:** Lane and Georgiev explore the potential of deep learning to revolutionize mobile sensing applications, including speech emotion recognition. The study investigates the feasibility of leveraging deep learning techniques for real-time emotion detection on mobile devices.

[5] Rázuri et al. (2015)**:** Rázuri et al. investigate speech emotion recognition within the context of emotional feedback for human-robot interaction. The paper explores the role of emotional cues in enhancing user experience and interaction satisfaction in human-robot interaction scenarios.

[6] Le and Provost (2013)**:** Le and Provost present a novel approach to emotion recognition from spontaneous speech using hidden Markov models with deep belief networks. The study proposes a hybrid model capable of capturing complex emotional states from unscripted speech data.

[7] Nassif et al. (2019)**:** Nassif et al. conduct a systematic review evaluating the effectiveness of speech recognition using deep neural networks. The paper provides insights into the current state-of-the-art techniques and their applicability in various speech recognition tasks.

[8] Lalitha et al. (2014)**:** Lalitha et al. contribute to the literature with a study on speech emotion recognition, focusing on algorithmic approaches and their practical applications. The paper discusses the design and implementation of speech emotion recognition systems, highlighting their potential impact on real-world applications.

[9] Scherer (2005)**:** Scherer provides foundational insights into the nature of emotions and their measurement. The paper discusses theoretical frameworks for understanding emotions and proposes methodologies for accurately measuring emotional states.

[10] Balomenos et al. (2004)**:** Balomenos et al. discuss emotion analysis in man-machine interaction systems, emphasizing the importance of emotion-aware interfaces for enhancing user engagement and satisfaction. The study explores techniques for detecting and interpreting emotional cues in human-computer interaction scenarios.

[11] Cowie et al. (2001):

Cowie et al. discuss emotion recognition in human-computer interaction, focusing on methodologies and techniques for accurately detecting and interpreting emotional cues. The paper explores the role of emotion recognition systems in enhancing user experience and interaction satisfaction in various human-computer interaction scenarios.

[12] Kwon et al. (2003):

Kwon et al. present a study on emotion recognition using speech signals, particularly focusing on the utilization of speech-based features for emotion classification. The paper discusses the design and implementation of a system capable of accurately recognizing emotions from speech data, presenting findings from empirical evaluations conducted at the EUROSPEECH conference.

[13] Picard (1995):

Picard introduces the concept of affective computing, laying the foundation for research at the intersection of computing and human emotions. The paper discusses the development of computational models capable of recognizing, interpreting, and responding to human emotions, paving the way for advancements in emotion recognition technology.

[14] Koolagudi and Rao (2012):

Koolagudi and Rao provide a comprehensive review of emotion recognition from speech, highlighting key methodologies, techniques, and challenges in the field. The paper discusses feature extraction, classification schemes, and available databases used in speech emotion recognition, offering insights into the state-of-the-art approaches.

[15] El Ayadi et al. (2011):

El Ayadi, Kamel, and Karray conduct a survey on speech emotion recognition, focusing on features, classification schemes, and databases utilized in the field. The paper provides a comprehensive overview of existing approaches to speech emotion recognition, highlighting the importance of feature selection, classification algorithms, and dataset availability in developing accurate emotion recognition systems.

Future Scope

Future Scope of Emotional Recognition in Speech Using Deep Learning:

1. Multimodal Emotion Recognition: Future research could explore the integration of multiple modalities, such as facial expressions, physiological signals, and textual cues, alongside speech data for more robust emotion recognition systems.

2. Contextual Understanding: Deep learning models can be further enhanced to understand contextual cues and situational factors that influence emotional expression in speech, leading to more context-aware emotion recognition systems.

3. Personalized Emotion Recognition: Tailoring emotion recognition systems to individual users by leveraging personalized training data and adaptive algorithms can improve the accuracy and effectiveness of emotion detection in various applications.

4. Real-Time Emotion Detection: Advancements in hardware and algorithm optimization can enable real-time emotion detection systems, facilitating applications such as emotion-aware virtual assistants, emotion-sensitive tutoring systems, and emotion-based feedback mechanisms.

5. Cross-Cultural Emotion Recognition: Research focusing on cross-cultural differences in emotional expression and perception can lead to more culturally sensitive emotion recognition models capable of accurately capturing and interpreting emotions across diverse populations.

6. Emotion Generation and Synthesis: Deep learning techniques can be employed not only for emotion recognition but also for generating and synthesizing emotional speech, enabling applications such as emotionally expressive virtual avatars and emotion-driven storytelling systems.

7. Ethical Considerations and Bias Mitigation: Future research should address ethical concerns related to privacy, data security, and potential biases in emotion recognition algorithms, ensuring fair and responsible deployment of these technologies in real-world settings.

8. Human-Robot Interaction: Deep learning-based emotion recognition can play a crucial role in improving human-robot interaction by enabling robots to perceive and respond to human emotions effectively, leading to more intuitive and empathetic robotic companions.

9. Healthcare and Wellbeing Applications: Emotion recognition systems can be applied in healthcare settings for early detection of mental health disorders, monitoring emotional states during therapy sessions, and providing personalized emotional support to individuals in need.

10. Continuous Learning and Adaptation: Developing adaptive deep learning models capable of continuous learning and adaptation to changing environments and user preferences can enhance the long-term performance and usability of emotion recognition systems in dynamic real-world scenarios.

By exploring these future directions, researchers can contribute to the development of more sophisticated, accurate, and socially aware emotion recognition systems that have the potential to positively impact various aspects of human life and interaction.

Conclusion:

In conclusion, the utilization of deep learning techniques for emotional recognition in speech marks a significant advancement in the field of human-computer interaction and affective computing. Through the exploration of various deep learning architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoders, researchers have made notable progress in accurately detecting and interpreting emotional cues from speech data. These advancements have paved the way for the development of emotion-aware systems capable of understanding and responding to human emotions in real-time, with applications spanning from virtual assistants to healthcare and robotics.

However, despite the promising results achieved thus far, several challenges and opportunities for future research remain. The integration of multimodal data sources, including facial expressions and physiological signals, could enhance the robustness and accuracy of emotion recognition systems. Additionally, addressing ethical considerations such as privacy, bias, and data security is paramount to ensuring the responsible deployment of these technologies in diverse settings.

Moreover, ongoing efforts to improve the contextual understanding of emotions in speech, personalize recognition systems, and enable real-time emotion detection are crucial for advancing the state-of-the-art in emotional recognition technology. By embracing these challenges and opportunities, researchers can continue to push the boundaries of emotion recognition in speech using deep learning, ultimately enhancing human-computer interaction, and fostering more empathetic and responsive systems in the digital age.

Github link:

Abhinavchoudhary007/DeepLearning: project of deep learning : Emotion recognition in Speech using Deep learning (github.com)

Reference:

1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Commun. ACM, vol. 61, no. 5, pp. 90–99, 2018.

[2] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," Inf. Fusion, vol. 49, pp. 69–78, Sep. 2019.

[3] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," IEEE Access, vol. 5, pp. 326–337, 2016.

[4] N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?" in Proc. ACM 16th Int. Workshop Mobile Comput. Syst. Appl., 2015, pp. 117–122.

[5] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, "Speech emotion recognition in emotional feedbackfor human-robot interaction," Int. J. Adv. Res. Artif. Intell., vol. 4, no. 2, pp. 20–27, 2015.

[6] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden MARKOV models with deep belief networks," in Proc. IEEE Workshop Autom. Speech Recognit. Understand., Dec. 2013, pp. 216–221.

[7] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," IEEE Access, vol. 7, pp. 19143–19165, 2019. [8] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in Proc. Int. Conf. Adv. Electron. Comput. Commun. (ICAECC), Oct. 2014, pp. 1–4.

[9] K. R. Scherer, “What are emotions? And how can they be measured?'' Social Sci. Inf., vol. 44, no. 4, pp. 695–729, 2005.

[10] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, “Emotion analysis in man-machine interaction systems,'' in Proc. Int. Workshop Mach. Learn. Multimodal Interact. Springer, 2004, pp. 318–328.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction,'' IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32–80, Jan. 2001.

[12] O. Kwon, K. Chan, J. Hao, T. Lee, “Emotion recognition by speech signal,'' in Proc. EUROSPEECH, Geneva, Switzerland, 2003, pp. 125–128.

[13] R. W. Picard, “Affective computing,'' Perceptual Comput. Sect., Media Lab., MIT, Cambridge, MA, USA, Tech. Rep., 1995.

[14] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: A review,'' Int. J. speech Technol., vol. 15, no. 2, pp. 99–117, 2012.

[15] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,'' Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.