BUBT | BANGLADESH UNIVERSITY OF BUSINESS AND TECHNOLOGY

Committed to Academic Excellence

Assignment-1

Course Title : Pattern Recognition

Course Code : CSE 467

Submitted by	Submitted to
Name : MD. Khaled Mahmud ID : 19202103198 Program: CSE Intake : 44 Section : 03	Tahiya Ahmed Chowdhury Lecturer Dept. of CSE BUBT

Submission Date: 01-12-2023

A Comprehensive Review of Speech Emotion Recognition Systems

Introduction: Speech Emotion Recognition (SER) is all about computers understanding feelings from how people speak. It helps them figure out if someone sounds happy, sad, angry, or another emotion. This tech uses computers to analyze things like pitch and tone in speech to tell what emotion someone is expressing.

Functions of SER:

Speech Emotion Recognition (SER) is a technology that enables machines to detect and analyze emotions from spoken language. It has various applications, including enhancing human-computer interaction, aiding virtual assistants and chatbots in understanding user input's emotional context, and assisting in customer service and feedback analysis by analyzing phone calls or other communication channels.

Background:

Paul Ekman, a psychologist, made a significant contribution to facial expression study in the 1960s by identifying six basic emotions. Paralinguistic study focused on nonverbal speech components, such as intonation, pitch, and rhythm, to identify emotions. Automated speech recognition (SER) emerged in the 1980s and 1990s, using acoustic features to classify emotions. AI and machine learning in the late 20th century transformed SER research, enabling more complex speech and emotion recognition systems.

Motivation:

Speech emotion recognition (SER) aims to enhance customer service, personalize education, detect mental health issues early, enhance entertainment, security, patient care, market insights, and facilitate natural human-robot interactions.

2. Literature Review:

SER uses various databases for training classification models, including CNNs for unsupervised feature extraction and transfer learning between ASR and SER databases. CNNs with MFSCs outperform previous implementations, while RNNs are used for utterance level classification. IEMOCAP database evaluates CNNs and LSTMs, while deep learning techniques with Mel scale log spectral coefficients are evaluated. Multimodal approaches combining audio and textual information are explored.

3. State of Art

● Trending models: Transformer-based models, such as BERT and GPT-3

● Trending dataset: RAVDESS, EMO-DB and IEMOCAP

● Trending feature extraction techniques: Mel-frequency cepstral coefficients

(MFCCs), Log-mel spectrograms, and Prosodic features

4.Feature extraction

The process involves utilizing Mel-frequency Cepstral Coefficients (MFCCs), Mel-scaled spectrograms, Chromagrams, Spectral contrast features, and Tonnetz representation for feature extraction.

5. Dataset

The "Continuous Speech Emotion Recognition with Convolutional Neural Networks" paper uses the "Acted Emotional Speech Dynamic Database" (AESDD) dataset for training and testing the model. AESDD is a database containing recordings of acted emotional speech utterances in Greek, showcasing five emotions: anger, disgust, fear, happiness, and sadness. The dataset is designed to grow dynamically through contributions from professionals in engineering, media, and performance fields. It features professional recordings, involving only professional actors, and being evaluated by a theatrologist to ensure the authenticity of emotional expressions.

6. Methodology

The study explores the use of Convolutional Neural Networks (CNNs) architecture, data augmentation techniques, and subjective evaluation of the Acted Emotional Speech Dynamic Database (AESDD).

7. Result

The CNN architecture outperforms baseline models by 8.4% in accuracy. Disgust is the most difficult emotion to distinguish. Female participants scored 74.2, while males scored 73.64. Participants from theatrical studies scored 74.9, while mass media and journalism scored 73.44. Subjective evaluation experiments validate the AESDD's suitability for SER purposes. A statistical relationship exists between participants' responses based on their academic studies.

8. Conclusion

The proposed CNN architecture outperforms SVM models by 8.4% in accuracy, and data augmentation doesn't affect classification accuracy in validation tests. Unsupervised feature extraction enables real-time systems. Future plans include language and speaker-dependent approaches.