**INTRODUCTION**

Speech Emotion Recognition (SER) can be viewed as a subfield of Automated Emotion Recognition (AER) because it uses the same type of signal, feature extraction procedures, and potential use of various machine learning techniques, including deep learning (DL) architectures, that are also used in Natural Language Processing (NLP). SER and ASR both share the sequential nature of the data. SER research has easily adopted feature extraction works that were previously created for ASR, such as the use of Mel Frequency Cepstral Coefficients (MFCC) for pattern identification and classification. Additionally, DL techniques like various types of Convolutional Neural Networks  (CNN) and Recurrent Neural Networks (RNN) that were created in ASR or NLP.[1]

**Modeling Emotions for Machine Learning:**

To be able to classify emotions using computer algorithms, we need to have a mathematical model describing them. The classical approach defined by psychologists is based on three measures that create a three-dimensional space that describes all the emotions. These measures or dimensions are pleasure, arousal, and dominance . A combination of these qualities will create a vector that will be in one of the defined emotion territories, and based on that, we can report the most relevant emotion.Using pleasure, arousal, and dominance, we can describe almost any emotion, but such a deterministic system will be very complex to implement for machine learning. Therefore, in machine learning studies, typically, we use statistical models and cluster samples into one of the named qualitative emotions such as anger, happiness, sadness, and so forth. To be able to classify and cluster any of the mentioned emotions, we need to model them using Sensors 2021, 21, 1249 3 of 27 features extracted from the speech; this is usually done by extracting different categories of prosody, voice quality, and spectral features. [2]

**Exploring Speech Emotion Recognition:**

Humans naturally communicate with one another through speech. It gives details on the speaker's sentiments, ideas, and moods as well as the communication setting.Numerous examples of affective computing systems that rely on machine learning or deep learning models are found in the literature. These systems include those that use images , audio, video , text , EEG, physiological signals .Speech Emotion Recognition (SER) has been a popular area of study with a wide range of applications in several fields. Among its many applications are lie detection and criminal investigations , medical diagnosis and monitoring robotic emotion expressions, machine-human interaction systems , call centre answering , robotic assistance and helpline systems , theatre performance and interaction enhancements , mental health and fitness analysis in the classroom and online teaching , emotional state recognition of drivers , and intelligence assistance ,  including digital advertisements, online gaming, and customer feedback evaluation.[3]

**Integrating Neuroanatomy into Speech Emotion Recognition:**

The human brain’s perception of an emotion is related to multiple parts of the limbic system, which indicates that the human brain’s emotional perception network may have a certain structure, and the structure causes the relevant parts of the human brain to be more sensitive to certain emotions. Could this structure be introduced into speech emotion recognition? Many parts of the human brain are sensitive to the same emotion, but there are differences in sensitivity. So, are there similar but not identical internal structures in the various part of the human brain.Different emotions are involved in different parts. For example, both anger and sadness are linked to the amygdala, but sadness is linked to the left thalamus, whereas anger is not. This means that the human brain has differences in the perception of different emotions and also shows that the differences in the perception of different parts of the human brain are related to the internal structure of the parts. [4]

**Enhancing Doctor-Patient Communication:**

         The study indicates that doctor-patient connections are crucial for  compliance and happiness with healthcare services. Physician trust and sense of expertise increased as a result of a fulfilling cognitive encounter. Masks, however, conceal a portion of our most important means of demonstrating empathy in specific circumstances, such as pandemics. Using vocals to convey empathy becomes essential. Numerous research have already demonstrated the ability of artificial neural networks to predict emotion categories across a range of nonverbal metrics. One of the key technologies for assessing the quality of healthcare service consultations is automatic speech emotion recognition (SER), which has been developed for many years in a variety of life contexts.[8]

**Defining Emotion in Speech Emotion Recognition:**

To successfully implement a speech emotion recognition system, we need to define and model emotion carefully. However, there is no consensus about the definition of emotion, and it is still an open problem in psychology. According to Plutchik, more than ninety definitions of emotion were proposed in the twentieth century . Emotions are convoluted psychological states that are composed of several components such as personal experience, physiological, behavioral, and communicative reactions. Based on these definitions, two models have become common in speech emotion recognition: discrete emotional model, and dimensional emotional model.Discrete emotion theory is based on the six categories of basic emotions; sadness, happiness, fear, anger, disgust, and surprise, as described by. These inborn and culturally independent emotions are experienced for a short period . Other emotions are obtained by the combination of the basic ones. Most of the existing SER systems focus on these basic [emotional categories](https://www.sciencedirect.com/topics/computer-science/emotional-category). In daily life, people use this model to define their observed emotions, hence labeling scheme based on emotional categories are intuitive. Nonetheless, these discrete categories of emotions are not able to define some of the complex [emotional states](https://www.sciencedirect.com/topics/computer-science/emotional-state) observed in daily communication.[6]

**Advancements in Speech Emotion Recognition:**

Speech is the most direct and natural method of communication between humans, and even between human and machine. However, we still cannot achieve natural interaction between humans and machines because the current machines cannot sufficiently understand the emotional status of humans. This discrepancy has led to a new research field: speech emotion recognition . In this paper, we focus on the algorithm analyzing speech characteristics and apply machine learning (ML) algorithm to recognize underlying emotions. Generally speaking, according to the emotion conceptualization, speech emotion recognition systems can be divided into two categories: continuous-label emotion recognition , and discrete-label emotion recognition. Early research on discrete speech emotion focused mainly on selecting speech The associate editor coordinating the review of this manuscript and approving it for publication was Alma Y. Alanis . acoustic features that can represent different emotions. Thus, many short- and long-term acoustic features combining static mathematics formulas have been proposed. The most popular approach is to extract a large number of features based on mathematic models at the utterance level. Then apply a statistical machine learning algorithm.[7]

**Advancing Human-Machine Interaction:**

The speech signal is the fastest and the most natural method of communication between humans. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine. However, this requires that the machine should have the sufficient intelligence to *recognize* human voices. Since the late fifties, there has been tremendous research on speech recognition, which refers to the process of converting the human speech into a sequence of words. However, despite the great progress made in speech recognition, we are still far from having a *natural* interaction between man and machine because the machine does not understand the *emotional* state of the speaker. This has introduced a relatively recent research field, namely speech emotion recognition, which is defined as extracting the emotional state of a speaker from his or her speech. It is believed that speech emotion recognition can be used to extract useful semantics from speech, and hence, improves the performance of speech recognition systems.Speech emotion recognition is particularly useful for applications which require natural man–machine interaction such as web movies and computer tutorial applications where the response of those systems to the user depends on the detected emotion . It is also useful for in-car board system where information of the mental state of the driver may be provided to the system to initiate his/her safety . It can be also employed as a diagnostic tool for therapists. It may be also useful in automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties. In aircraft cockpits, it has been found that speech recognition systems trained to stressed-speech achieve better performance than those trained by normal speech. Speech emotion recognition has also been used in call center applications and mobile communication. The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice.[5]

**Challenges in Speech Emotion Identification:**

               The most popular and efficient mode of communication is through speech, and one of the many intricate functions of the human brain is speech comprehension.Many different types of information, such as gender, words, accent, mood, and age, can be extracted from a speech signal and used in a variety of ways. Speech emotion identification is one of the hardest things for researchers to do in speech processing.[9]

**References**

[1]   de Lope, Javier, and Manuel Graña. "An ongoing review of speech emotion recognition."    *Neurocomputing* 528 (2023): 1-11.

[2] Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep   learning techniques for speech emotion recognition, from databases to models." *Sensors* 21, no. 4 (2021): 1249.

[3] George, Swapna Mol, and P. Muhamed Ilyas. "A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise." *Neurocomputing* (2023): 127015.

[4] Liu, Gang, Shifang Cai, and Ce Wang. "Speech emotion recognition based on emotion perception." *EURASIP Journal on Audio, Speech, and Music Processing* 2023, no. 1 (2023): 22.

[5] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern recognition* 44, no. 3 (2011): 572-587.

[6] Akçay, Mehmet Berkehan, and Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers." *Speech Communication* 116 (2020): 56-76.

[7]  Sun, Ting-Wei. "End-to-end speech emotion recognition with gender information." *IEEE Access* 8 (2020): 152423-152438.

[8] Huang, Ching-Tzu, Chih-Wei Huang, Hsuan-Chia Yang, and Yu-Chuan Jack Li. "Speech Emotion Recognition Applied to Real-World Medical Consultation." In *MEDINFO 2023—The Future Is Accessible*, pp. 1121-1125. IOS Press, 2024.

[9] Wani, Taiba Majid, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. "A comprehensive review of speech emotion recognition systems." *IEEE access* 9 (2021): 47795-47814.