Speech Emotion Recognition

Speech is the main and most effective way of human communication. There is no transfer of emotion just by speaking the words in a straight tone. There is a transfer of emotion in every person's speech, even if he is not aware of it. Detecting this emotional state is also a difficult task, because when a person says "awesome" they may be making an allusion to it. But it is a fact that there are physiological changes in people according to their emotional state. There are also devices that determine mood by transferring data from things such as heart rate, blood pressure, blood pressure to a device. But without the need for a device, it is possible to determine the mood from facial expressions and voice. There are also an increase in studies that describe emotion from face and voice.

It is important in the process of recruiting people, in the process of assessing human emotions, in matters such as detecting lies and other issues like this. There are many ways in which the various physical properties of humans can be understood by machines. Among these, facial expressions can be understood by eye movements or facial recognition systems. Or their body language and gestures are among them. In addition to these, besides the physical movements that people have, speaking is the most effective and easy way of communication. Most of the mood analysis studies and articles are made on the act of speaking that people have rather than physical movements.

In a few articles reviewed and made, there are applications such as changing the presentation style of the instructor by evaluating the situation of the students being uninterested or bored during the education in online education and applications, thanks to the warning received by the system. At the same time, in another study conducted in the field of medicine, it was ensured that the patient's health status was examined by the patient's voice according to the patient's speech. In another way, work has been done on robots that are important for the future. In this study, robots were trained according to human emotions. The basis of this is to provide the most appropriate communication between human and robot. It has a wide and comprehensive application area, which detects the behavior and emotional states of the people on the other side and enables them to react accordingly, in calls with call centers or in another study included with these.
Education, automobile, security, communication, health can be given as examples as the field of systems made in this field.[1]

Feature Extraction

Feature extraction plays a crucial role in the speech recognition process and extracting valuable data from sample speech has been a crucial part of research for many years. Feature extraction helps distinguish one speech from another. It converts the raw speech signal into a dense but effective representation that is more stable than the original, from which it is possible to reconstruct it from the original signal.

Speech features can be grouped into 4 main categories. These are: continuous features, qualitative features, spectral features, and TEO (Teager energy operator)-based features. This section will examine the pros and cons of each category. In addition, speech emotion recognition can be performed by combining features belonging to different categories for speech signal. The following figure shows examples of properties for each category.

Continuous Speech Features

Most researchers believe that continuous features such as pitch and energy's express most of the emotion in the voice.[2,3,4] According to studies by Williams and Stevens [5], low activation versus high activation affects overall energy, and energy distribution affects the duration of the pause frequency. This result has been confirmed by other studies.[6,7] Acoustic connections related to continuous sound are given below.

(1) pitch-related features

(2) formants features

(3) energy-related features

(4) timing features

(5) articulation features

Voice Quality Features

Experiments on humans have proven that the quality of a voice plays an important role in conveying emotion. [4,8,9,10] The sound quality seems to complement fully developed emotions in the most regular way, that is, it can lead people to action. [4] Acoustic correlations for sound quality are given below.

(1) voice level: signal amplitude, energy and duration have been shown to be reliable measures of voice level;

(2) voice pitch;

(3) phrase, phoneme, word and feature boundaries

(4) temporal structures

Spectral-Based Speech Features

Spectral features are often chosen as a short-term representation for the speech signal. It is accepted that the emotional orientation of a sound has an effect on the distribution of spectral energy throughout the speech frequency range.[11] For example, it has been proven by studies that expressions containing happiness have high energy in the high frequency range, and expressions containing sadness have low energy in the same range.[12,13]

Nonlinear TEO-Based Features

According to studies by Teager, speech is produced by the non-linear airflow in the sound system.In stressful conditions, the air flow in the system is affected by the muscle tension of the person producing the sound. For this reason, non-linear speech features are required to perceive the emotion in the voice. It has been observed that the fundamental frequency changes under stressful conditions, as is the distribution of harmonics over the critical bands.[14,15] It has been verified that the TEO of the multiple frequency signal reflects not only the individual frequency but also the interaction between them.[16] Relying on this validation, TEO-based features can be used to find stress in speech.

ML, DNN and Dataset

An Urdu Speech Corpus for Emotion Recognition [17]. An emotional speech corpus for the Urdu language was designed and developed in this study. Different machine-learning techniques were used to identify emotions from Urdu speech signals. Five Urdu sentences—one each for happy, sad, angry, disgusted, and neutral—were simulated into five different emotional states. The highest overall recognition accuracy for the disgust emotion for males, females, and the entire dataset was 72.5% with "k-NN," 68.5% with "one-against-rest classifier," and 66.2% with "k-NN." The maximum highest recognition accuracy for the dataset without the disgust emotion was 82.5% with "k-NN," 78.5% with "one-against-rest classifier," and 76.5% with "k-NN" for male, female, and the whole dataset, respectively.

Speech Emotion Recognition with Deep Learning [18]. In this paper, the Ryerson Multimedia Lab (RML) emotion database was used, which comprises 241 audiovisual and emotional expression examples in English. There are six fundamental human emotions: anger, disgust, fear, happiness, sadness, and surprise. Began adjusting the basic AE's settings to improve identification rates when the RBF kernel of the SVM classifier was used. We acquire a superior recognition rate equal to 70.37% when the number of units in the hidden layer is 35 after a series of trials in which the number of units is varied.

Leila et al [19]. In this paper, described an automatic speech emotion recognition (SER) system that classifies seven emotions using three machine learning methods (MLR, SVM, and RNN). In order to offer a mix of these properties, two types of features (MFCC and MS) were derived from two separate acted databases (Berlin and Spanish databases). The machine learning models were trained and tested for their ability to distinguish emotional states based on these characteristics. When speaker normalization (SN) and feature selection (FS) are applied to the features, according to SER, the recognition rate of the Berlin database is 83% accurate across the board. This results showed that RNN frequently performs better with more data and that it has the constraint of having very long training cycles. As a result, we came to the conclusion that, as compared to RNN, the SVM and MLR models had a better chance of being used practically with little amounts of data.

Multimodal Speech Emotion Recognition and Ambiguity Resolution [20]. In this paper, the IEMOCAP database was used. This dataset was used firstly only for audio tests, then only for text tests and finally by combining audio + text for 6 emotions(Angry, Happy, Sad, Fear, Surprise, Neutral). This dataset is used for RF, XGB, SVM, MNB, LR, MLP, LSTM, ARE and E1. Here only the highest accuracy for Audio was achieved with E1 (Ensemble (RF + XGB + MLP)) at 56.6. For Text only, the highest accuracy was achieved with the TRE (Text-Recurrent Encoders) as 65.5. In the last step, when working on the model again by setting Audio+Text, the highest accuracy was obtained with MDRE (Multimodal Dual-Recurrent Encoders) of 75.3.

Clustering-Based Speech Emotion Recognition (2020) [21]. In this paper, this research outlined an innovative strategy for SER to enhance the Boost recognition accuracy while cutting down on model costs and processing times. On the other hand, we proposed a new method to choose a more effective speech sequence using the K-mean clustering approach based on RBF, and changing By using the STFT technique, it may be turned into spectrograms. Hence, We identified the prominent and discriminative characteristics from voice signal spectrograms by using the "FC-1000" Resnet layers in the CNN model are added before it is normalized. Using mean and standard deviation, we can eliminate the variation. After normalization, we feed these discriminative features to deep BiLSTM to learn the hidden information, recognize the final state of the sequence, and classify the emotional state of the speakers. To test the system's resilience, we assessed it using three common datasets: IEMOCAP, EMO-DB, and RAVDESS. We increase the recognition accuracy for the IEMOCAP dataset to 72.25%, the EMO-DB dataset to 85.57%, and the RAVDESS dataset to 77.02%.

Speech emotion recognition with deep convolutional neural networks [22]. In this pape, the IEMOCAP, EMO-DB, and RAVDEES databases were used in this paper. Afterward, we used an incremental approach to enhance classification accuracy by tweaking our baseline model. Unlike some earlier methods, none of the suggested models require translation to visual representations before working with raw audio data. Our best-performing model exceeds previous frameworks for RAVDESS and IEMOCAP, establishing a new state-of-the-art. It performs better than all prior efforts for the EMO-DB dataset, with the exception of one, and compares favorably with that one in terms of generality, simplicity, and application. Specifically, more precise, the suggested framework completes speaker-independent audio classification tasks with 71.61% for RAVDESS with 8 classes, 86.1% for EMO-DB with 535 samples in 7 classes, 95.71% for EMO-DB with 520 samples in 7 classes, and 64.3% for IEMOCAP with 4 classes.

Turkish and Other Languages

Papers	Dataset	Emotions	Technique	Accuracy (%)
An Urdu Speech Corpus For Emotion Recognition (2022) [17]	Urdu Emotional Speech Dataset	Angry, Happy, Sad, Neutral	k-NN (with disgust) k-NN (without disgust)	72.5 82.5
Clustering-Based Speech Emotion Recognition (2020) [21]	IEMOCAP EMO-DB RAVDEES	Angry, Happy, Sad, Fear, Surprise, Neutral	CNN + LSTM	72.25 85.57 77.02
Speech Emotion Recognition with Deep Learning (2020) [18]	RML Dataset	Angry, Disgust, Fear, Happy, Sad, Surprise	Basic AE with SVM Stacked AE with SVM	72.83 74.07
Speech emotion recognition with deep	IEMOCAP EMO-DB RAVDEES	Angry, Disgust, Fear, Happy, Sad, Surprise	CNN LSTM	64.30 71.61 86.1

convolutional neural networks (2020) [22]				
Multimodal Speech Emotion Recognition and Ambiguity Resolution (2019) [20]	IEMOCAP (Audio Only)	Angry, Happy, Sad, Fear, Surprise, Neutral	RF XGB SVM MNB LR MLP LSTM ARE (4-class) E1 (4 -class) E1	56.0 56.6 33.7 31.3 33.4 41.0 43.6 56.0 56.3 56.6
Multimodal Speech Emotion Recognition and Ambiguity Resolution (2019) [20]	IEMOCAP (Text Only)	Angry, Happy, Sad, Fear, Surprise, Neutral	RF XGB SVM MNB LR MLP LSTM TRE (4-class) E1 (4 -class) E1	62.2 56.9 62.1 61.9 64.2 60.6 63.1 65.5 63.1 64.9
Multimodal Speech Emotion Recognition and Ambiguity Resolution (2019) [20]	IEMOCAP (Audio+Text)	Angry, Happy, Sad, Fear, Surprise, Neutral	RF XGB SVM MNB LR MLP LSTM ARE (4-class) E1 (4 -class) E1	65.3 62.2 63.4 60.5 66.1 63.2 64.2 75.3 70.3 70.1
Leila et al. (2019) [19]	Berlin Emo DB	Angry, Disgust, Joy, Neutral Happy, Sad, Surprise	SVM, MLR	83.0

Speech To Text - Audio Recognition

Neural networks [23] play a major role in the use and definition of human behaviour and movements in large areas such as artificial intelligence, machine learning, and deep learning. In this study, we will try to describe the behaviors coming from hostile environments with the information we get from neural networks. While studies on speech recognition are the most common in the world, studies on voice are less. In automatic speech, a sound wave is sent, and an inference is made about what language and how it is and converted into text. In this article, we will examine the study of text from a speech on a transcription neural network.

Several suitable datasets are used to investigate the effects of neural networks on speech-to-text. The Mozilla Common Voice dataset [24] is made available to evaluate the effectiveness of attacks made in this article. In the study, randomly sent data (a) was used to find the wrong ones, while (b) was used to find the correct ones. In addition to this, another data set usage was presented to the project. In this dataset, the input waveform is converted to sound 50 times per second and gives the character probability distribution in this project. That means placing it at 50 characters per second. Thus, it is possible to convert short audio clips to text. In such an attack, it is effective even though it requires an average of 18dB distortion. Another dataset, it is concerned with finding the target expression of a random non-speech voice sample. In this study, there was no need to apply technical innovation, instead, only random sound waves were sent. The effectiveness of this sent sound wave was obtained by collecting five-second clips from classical music that does not contain any speech and target sentence parts in the common audio dataset. As a result of this study, it was observed that distortion up to a maximum of 20 dB occurred. In the last part of the study, data was obtained by providing silence. And no data set is used in this part. A hostile noise was added that caused the study not to write any type of text, and no results were obtained.

The thread model was used as a model in the study. In this model, given a sound wave sample and target y, the goal is to generate another sound wave. Here, if the output completely overlaps with the target without typos, the result is successful. Models and parameters from previous studies are defined as white boxes. it is a kind of competitor threat model. The threats to later work are the black box.

In the study, distortion metric is used to understand and solve a possible distortion. They measure distortion in decibels. It is intended that the disruption be minimal, and that people do not understand it.

Contributions

When we look at this paper, we see that there are many studies in the field of Speech Emotion Recognition and each study ends their work with a different technique. We aim to design a system that determines 4 different emotional states: Angry, Happy, Sad, Neutral. While detecting these emotions, we are planning a system that extracts emotion analysis from the voices of the speaker. In previous studies, similar systems were designed, and the accuracy rates of these systems were calculated by techniques as k-NN, CNN, RF, XGB, and SVM, and an average of 60-65 percent successes was determined. In these studies, datasets such as Urdu Emotional

Speech Dataset, RML Dataset, IEMOCAP or Text only were used as datasets. Although we plan to use 4 different situations as Angry, Happy, Sad, Neutral, different emotional analyses such as Fear, Disgust, Surprise, Joy was also included in these studies.

Although the English language dataset is used in most of these studies, many different languages, including Turkish, are used in a substantial part.

The main operation of the system we want to do is to first make an emotion analysis from the voice and then develop a system that makes speech to text and makes emotion analysis from that text. Then, to combine these two different analyses and produce a result.

References

[1] Sucksmith, E., Allison, C., Baron-Cohen, S., Chakrabarti, B., & Hoekstra, R. A. Empathy and emotion recognition in people with autism, first-degree relatives, and controls. Neuropsychologia, 51(1), 98-105,2013

[2] L. Bosch, Emotions, speech and the asr framework, Speech Commun. 40 (2003) 213–225.

[3] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE Trans. Audio Speech Language Process. 17 (4) (2009) 582–596.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (2001) 32–80.

[5] C. Williams, K. Stevens, Vocal correlates of emotional states, Speech Evaluation in Psychiatry, Grune and Stratton, 1981, pp. 189–220.

[6] T. Johnstone, K.R. Scherer, Vocal Communication of Emotion, second ed., Guilford, New York, 2000, pp. 226–235.

[7] R. Cowie, R.R. Cornelius, Describing the emotional states that are expressed in speech, Speech Commun. 40 (1–2) (2003) 5–32.

[8] K.R. Scherer, Vocal affect expression. A review and a model for future research, Psychological Bull. 99 (2) (1986) 143–165 cited by (since 1996) 311.

[9] J.R. Davitz, The Communication of Emotional Meaning, McGraw-Hill, New York, 1964.

[17] C. Gobl, A.N. Chasaide, The role of voice quality in communicating emotion, mood and attitude, Speech Commun. 40 (1–2) (2003) 189–212.

[11] T. Nwe, S. Foo, L. De Silva, Speech emotion recognition using hidden Markov models, Speech Commun. 41 (2003) 603–623.

[12] R. Banse, K. Scherer, Acoustic profiles in vocal emotion expression, J. Pers. Soc. Psychol. 70 (3) (1996) 614–636.

[13] L. Kaiser, Communication of affects by single vowels, Synthese 14 (4) (1962) 300–319.

[14] S. Bou-Ghazale, J. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Trans. Speech Audio Process. 8 (4) (2000) 429–442.

[15] H. Teager, S. Teager, Evidence for nonlinear production mechanisms in the vocal tract, in: Speech Production and Speech Modelling, Nato Advanced Institute, vol. 55, 1990, pp. 241–261.

[16] G. Zhou, J. Hansen, J. Kaiser, Nonlinear feature based classification of speech under stress, IEEE Trans. Speech Audio Process. 9 (3) (2001) 201–216.

[17] Asghar, A., Sohaib, S., Iftikhar, S., Shafi, M., & Fatima, K. (2022). An Urdu speech corpus for emotion recognition. PeerJ Computer Science, 8, e954.

[18] Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. Procedia Computer Science, 176, 251-260.

[19] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using machine learning. In Social media and machine learning. IntechOpen.

[20] Sahu, G. (2019). Multimodal speech emotion recognition and ambiguity resolution. arXiv preprint arXiv:1904.06022.

[21] Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access, 8, 79861-79875.

[22] Harár, P., Burget, R., & Dutta, M. K. (2017, February). Speech emotion recognition with deep learning. In 2017 4th International conference on signal processing and integrated networks (SPIN) (pp. 137-140). IEEE.

[23] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connection- ´ ist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376. ACM, 2006.

[24] Mozilla. Project deepspeech. https://github.com/mozilla/DeepSpeech, 2017.