SPEECH EMOTION RECOGNITION USING MFCC FEATURES

Mrs.P.Jyothi[1]., B.Sujala[2]., D.Sri Lakshmi Devi[3]., K.Sony[4] ., M.Dharanika[5]

1 Assistant Professor, Department of ECE., Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India (✉ shivajyothi29@gmail.com)

2, 3, 4, 5 B.Tech ECE, (19RG1A0404, 19RG1A0411, 19RG1A0428, 19RG1A0432),

Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

Abstract

It is a method of identifying the mental state to be able to identify even the most fundamental of emotional expressions in a person's spoken words. Engineers in the field of pattern recognition are becoming more interested in the task of emotion detection via speech. The relevance of emotions in human existence cannot be overstated. It's a crucial channel for communicating ideas, emotions, and one's overall state of mind to other people. Humans are hardwired with the capacity to read emotions from a speaker's tone and inflection. In the last decade, research into emotional computing has exploded with the growth of Human Computer Interaction. Video surveillance, interactive entertainment, intelligent car systems, and medical diagnostics are all fields where an emotionally-aware computer system might improve communication with humans. In this study, we use Mel Frequency Campestral Coefficient features and Support Vector Machine classifiers to categories emotional states. The accuracy of feature recognition is assessed since it is similar to how humans hear. To see how these traits may be used for emotion recognition, consider the following examples.

INTRODUCTION

Human speech is both the quickest and most natural means of signaling to one another. Scientists have been inspired by this realization to consider spoken language as a potential medium for rapid and effective machine-human communication. In order for this to work, however, the computer will need to be intelligent enough to recognize human voices. Voice recognition, or the technology used to decode human speech into text, has been the subject of intensive study since the 1960s. There has been significant development in voice recognition, but this is still a long way from allowing for human and machine to connect in a natural way. Because of this, speech emotion recognition has emerged as a new area of study, which attempts to determine the speaker's emotional state by analyzing their words. It is hypothesized that by recognizing emotions in a speaker's voice, valuable semantics may be extracted

From speech, which in turn boosts performance? Emotions may be happy or negative and last for a very short amount of time.

Why is it required?

If the system can identify anger or irritation in the speaker's voice, it may modify its response to the user accordingly.

Emotion Speech Recognition is challenging task

The purpose of using Speech Emotion Recognition (SER) is to modify the system's reaction when it detects emotions like irritation or dissatisfaction in the speaker's voice. Emotion 2.2 Recognizing speech is a very difficult undertaking. 1. It is unclear whether aspects of speech are best at differentiating between emotions. 2. Another challenge is posed by the acoustic variability presented by the presence of diverse sentences, speakers, speaking styles, and speaking speeds, all of which directly impact most of the commonly retrieved speech variables, including pitch and energy contours. Third, we can tell how someone is feeling only by listening to their voice or looking at their face. In non-biological emotion detection systems, we aim to either identify the emotions by reading the subject's facial expressions or by the subject's voice, and there have even been efforts to categories emotionally expressive phrases and distinguish emotions from them. However, humans process emotional recognition in a fundamentally different manner. Human emotions evolved through time, and as such, distinct regions of the brain are responsible for handling the various feelings we experience. With the help of neural mapping and research into the limbic system, we now have a better understanding of how the brain's neuronal network interacts with neurotransmitters like dopamine and serotonin to identify and generate emotional reactions. A speaker's pitch, volume, and tempo of speech all change depending on the emotion being conveyed.

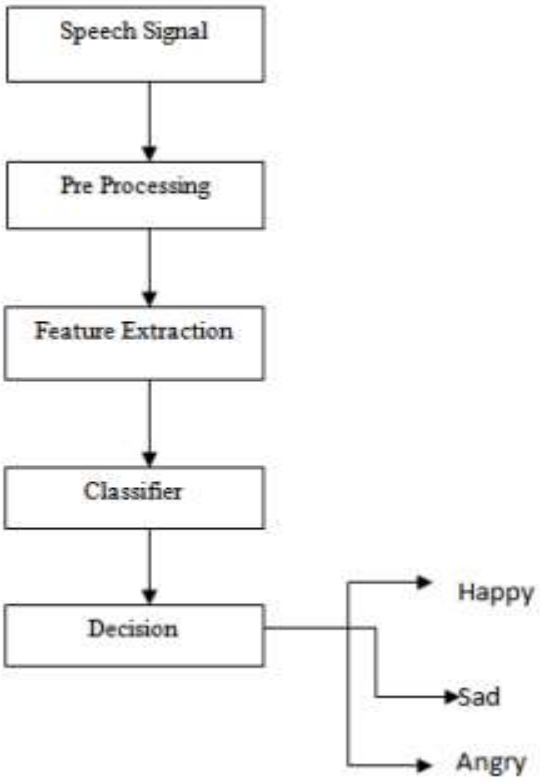

Fig 2.2.Flow chart of implementation of the proposed system.

Steps of proposed method

a) Pre processing
b) Framing
c) Windowing
d) FFT
e) Feature extraction
f) Classifier
a) Preprocessing

Preprocessing is the set of steps that must be taken on speech signal time samples before feature extraction may take place. For instance, every speech must undergo some kind of energy normalization in order to account for variations in recording environments. The length of the signal is reduced by cutting off the inaudible portions of the original speech, which add up to nothing. Energy levels of signals are calculated in order to make them comparable. Frames of varying durations are used to evaluate speech signals. First, the signal is demised by soft thresholding the coefficients; next, the quiet portions of the signal are deleted by thresholding the signal's energy, since they provide no information.

b) Framing;

After the voice signal has been pre-emphasized, it is partitioned into frames of N sample points, with M seconds between each pair of frames (lower than N). The first N sample points make up the first frame. Each successive frame overlaps the previous one by a factor of N-M sample points, thus the second frame begins at the Math sample point after the first frame. This procedure is repeated until everyone can fit into one or more of the available picture frames. Our experiments use a frame length of N = 256. (10ms). each frame overlaps the one next to it to guarantee smooth movement between them.

Windowing

Each frame is given a Hamming window to smooth out any blips in the signal and make sure that the initial and final data points are consistent with one another. The signal gaps at the beginning and end of each frame are reduced by windowing the frames.

d) FFT;

The frequency response of each frame is obtained by transforming the corresponding time domain signals into the frequency domain.

e) Feature extraction;

The process entails isolating relevant data from the provided speech and discarding any extraneous material. Energy, pitch, power, and MFCC are some of the features that may be retrieved. Pitch is the degree to which a tone is audible to the listener. Our sense of hearing is the bedrock of pitch. It's a feature of language that's readily apparent even to laypeople, and it's often misunderstood as the key to understanding sentiment. While an increase in pitch is often indicative of increased arousal, the shape of the pitch contour may also shed light on the speaker's emotional state. It is possible to determine pitch in either the time or frequency domain. Non-voiced components of a speech signal do not have pitch. Energy According to the human ear, loudness measures the intensity of a sound. Because direct measurement is challenging, the signal's energy is often employed as a proxy. After a Fourier transform is performed on the original signal, the energy may be determined from the spectrum. Like with pitch, a high energy level often indicates arousal, but subtle shifts in the energy curve might provide light on the speaker's state of mind.

MFCC Parametric representations of speech such as Mel-frequency campestral coefficients (MFCCs) are widely employed for automated speech recognition but have also shown promise in other areas, such as speaker identification and emotion detection. According to the developers, these are the most versatile qualities for use in any kind of speech assignment. The perceived pitch or frequency of a tone is measured in units called meals. One way in which MFCCs improve signal representation for human perception is by mapping onto the Mel scale, a frequency scale tailored to the human ear. When the Fourier transform of a windowed signal is applied to a Mel-scale filter bank, the results are obtained. In the next step, the logarithm zed spectrum is converted into a cestrum using a DCT (discrete cosine transforms). The cutoff frequencies of Mel filter banks are set by the midpoints of the two neighboring filters, which are triangular in shape. The filters feature midpoints that are equally spaced along the frequency axis and a constant band width measured in Mel. By using the logarithm, multiplication is converted to addition. It makes addition out of the FT's multiplication of magnitude.

F) SVM Classifier:

Emotion-related speech feature identification is a very difficult problem. Emotional states including anger, sorrow, fear, happiness, and boredom may all be categorized using a Support Vector Machine. When compared to other classifiers, SVM's classification performance stands out as particularly strong due to the algorithm's simplicity and efficiency. Statistical support vector machines (SVMs) are widely used for a variety of learning tasks, including classification, regression, and more. If you just have a few training samples to work with, SVM will perform better. However, we currently lack instructions for selecting an improved kernel with SVM parameter optimizations. There is no standard procedure for selecting an SVM and its parameters, or for selecting a kernel function and its parameters. Methods for determining the best settings for the SVM's parameters and kernel function were proposed in this study. This is how the system works: First, we'll isolate the features of utterances that correspond to different emotional states. The second stage of an improved method entails boosting the SVM's classification accuracy rate. The third step is the training of a classification model that takes use of the optimizations made in the previous two steps. Fourth, the system classifies the test samples and returns a result (class label or recognition rate). A key tenet of support vector machines is to use a hyper plane as the decision surface, which maximizes the gap between false-positive and false-negative examples. As a result, SVM is best suited for identifying two-class patterns. Combining binary support vector machines may help with many different pattern classification issues.

RESULTS:

Finally, a voice signal is categorized according to its predominant emotions, such as happiness, rage, or apathy. Specifically, a mat lab graphical user interface is used.

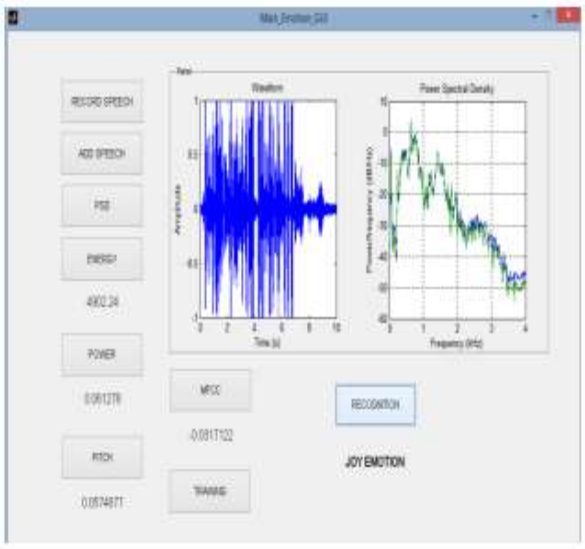

Fig4.1Matlab GUI for emotion recognition using speech

CONCLUSION

More and more people are curious in creating human-like robots as technology advances. The necessity of having happy users grows as more and more technological items enter the market. With the help of affective computing, we may finally have an interface that feels natural to use and adapts to the user's preferences. Emotions are the central focus of affective computing. Affective computing encompasses any studies that aim to detect, recognize, or create emotions. Any emotion detection technology would be able to gauge user happiness or discontent. Such systems might also be used to identify negative emotions, such rage or frustration, in addition to positive ones, like happiness. Similar to when one is behind the wheel of an automobile, the user may be restrained. Speech and facial emotion detections are the most common types of emotion detection tasks. Its widespread adoption may be attributed to the ease with which users can access their own face or voice data. There is a wealth of information contained inside human speech. Interpersonal discourse is the medium via which humans exchange information with one another. The acoustic component of human speech contains crucial information regarding affect. The features are extracted using MFCC. The SVM-based algorithm's overall performance is evaluated.

AKNOWLEDGEMENT

I'd want to take this opportunity to thank everyone who had a role in making it possible for me to finish the paper. Firstly, I'd want to give my thanks to Driven. Nit aware, who has been a great mentor to me. This project would not have been a success without his insightful advice. I'd like to thank Prof.Santhosh Bari, head of the Electronics and Communication Engineering department, for all the help and motivation he's given me. I'd want to express my gratitude to our college's principal, Driven. Nit aware, for ensuring that we have all we need to do our job.\

REFERENCES:

[1] Mehrdad J. Ganged, AliGhodsi, Mohamed S. Kamel,"Multiview Supervised Dictionary Learning in Speech Emotion Recognition," IEEE Transaction on audio, speech, and language processing.

[2] Sheikh Gupta1, Jafreezal Jaafar2, Wan Fatimah wan Ahmad3 and Armpit Bansal4 J. Clerk Maxwell, "Feature extraction using mock" Signal & Image Processing : An International Journal (SIPIJ) Vol.4, No.4, August 2013

[3] N.Murali Krishna1,P.V. Lakshmi2,Y. Srinivas3 J.Sirisha Devi4," Emotion Recognition using Dynamic Time Warping Technique for Isolated Words," IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011

[4] Aisha Joshi," Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Volume 3, Issue 8, August 2013.

[5] Slam Mansur mohammed1, Mohammed Sharif Sayed2," LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification ," International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 6, No. 3, June,2013

