ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Accuracy of speech emotion recognition through deep neural network and k-nearest

[1] R.B Pradeeba, [2] K.Tarunika, [3] Dr.P.Aruna
[1][2][3] Department of computing, Coimbatore Institute of Technology.

***Abstract* - Emotion is a positive or negative state of a person's mind which is related to physiological activities. The main objective of the paper is to apply Deep Neural Network (DNN) and k-nearest neighbor (k-NN) in recognition of emotion from speech-especially scary state of mind. The most precise model will be utilized to generate alert signals through the cloud. Health care system and in particular intensive and palliative care is supposed to be the area of application. Utterance level is the main feature for analysis with special emphasis. Raw data collection, conversion of acoustic signals to waveforms, probable emotion classification and recognition using databases existing, creating alert signals through the cloud is the sequence of steps followed. The finding of the paper is a fruitful contribution to the palliative care system.**

Keywords: Emotion, recognition, palliative, utterance level, analysis, deep neural network.

I. INTRODUCTION

In the recent past, great progress has been made in artificial intelligence but we are far behind from naturally interacting with machines. This is because; the speech emotion recognition remains a very challenging task.

It is well established that emotions cause mental and physiological changes that are reflected in uttered speech. Speech is the medium of communication to express the mood or condition or feelings of that individual. Mood is a generalized, internal state of feeling. Mood of the person constantly changes according to the environment and surrounding. A high level arousal is associated with the mood or emotions like joy, anger and surprise whereas, , a low level arousal is associated with boredom or sadness. Emotion content of speech does not depend on the speaker or the lexical content. Decrypting emotions through several features has been a challenging research issue and a rule to follow is not yet established.

The present study aims to identify the highly effective classifier for recognizing the emotion, particularly panicked state. This paper conveys the 2 phases of analysis

1) voice to text conversion analysis
2) mood analysis through utterance level

Both the analysis thrive it's application over the usage of 2 main techniques of machine learning, they are k-nearest neighbor and deep neural network. the widely used classifier namely the K-nearest neighbor (KNN) which has a proven record of 91% accuracy and a Deep Neural Network (DNN) which is capable of learning high-level representation from raw features and effectively classifying data. Speech emotion recognition system extraction of proper features of speech signal which represents emotions is an essential factor

This paper makes its view over the mood analysis from conversion of voice to text and testing the mood of a person by using the parameters of voice such as frequency, mel-frequency, pitch, Mel spectrogram, Harmonic percussive, Chromagram, Mel frequency cepstral, Beat tracking, Beat-synchronous features aggregation. The work proposed to carry out these two analysis and the experimental setup used for the following is discussed further. Both these analysis replay on two techniques called knn and dnn. Here the section[2] deals with the proposed work over the two ideologies. Section [3] contains the experimental setup of the phenomenon. Section [4] explains the further processes to be applied. Section [5] gives the further application and uses over this technique.

2. PROPOSED WORK:

Speech is vocalized form of communication used by humans, which explicitly expresses the mood or condition of the person during direct interaction. In case of indirect interaction machines are fed with the ideologies to analyze the mood of the person. As per the above mentioned 2 ideology. The voice is split and analysis is made over the techniques. Voice of the person depicts various pitch-frequency, which could be found through the variation in tone while speaking. These variations acts as a key factor of mood analysis in an indirect communication. The works proposed over these ideologies are:

2.1 Voice to text conversion analysis:

Voice received as input is first converted into textual form and displayed , the converted text is taken under analysis. Each word sample is made into consideration since each word of the combination bears its own related set of emotions within itself. On filtering the casual word on normal utterance, some words like abandoned, accepted, aggressive, alienated, amazed, amused, anxious, apathetic, astonished, aversion, awed, awful, bored, confident, confused, courageous, depression, critical, despair, detestable, devastate, disappointed, disillusioned, dismayed, distant,e cstatic, embarrassed, empty, enraged, frightened, fullfilled, ill, furious, guilty, hostile, ignored, inferior, infuriated, Inquisitive, insignificant, isolate, joy, judgemental, liberated, lonely, mad, ooptimistic, overwhelm, peaceful, perplexed, proud, ridicule, sarcastic, skeptical, submissive, terrified, vicitmized, vulnerable have perfect relationship over basic emotions like fear, anger, sadness, happiness, disgust, surprise. All these are maintained as database , comparisons and filtering is made to check the state of depression and to provide the right result as output

2.2 Mood analysis through utterance level:

Understanding the mood of the person in a direct conversation is just an identification, whereas the detection of mood in an indirect conversation is intelligence. For this intelligence machines require some parameters

Such as frequency, pulse, amplitude, structure, harmonic, pitch, mel-frequency.

Frequency: Variation in the pitch of voice

Pulse: Standard deviation in voice that indicates the rate of speaker

Amplitude: Variation in loudness of voice

Structure: Convey the voiced or unvoiced frame structure.

Harmonic: relative highness or lowness of voice

Pitch: conveys the mean of the voice and peaks of the sound spectrum of voice.

Tone of the voice bear all these parameters in such way that each emotion coordinate different ratio.

The parameter ratio of a person in:

Relaxed	:	stressed	=	0 : 1.5
Content	:	angry	=	-1.5 : 1
Bored	:	interested	=	-1.7 : 1.7
Sad	:	happy	=	-0.1 : 1.1
Friendly	:	hostile	=	-1.8 : 1.8
Timid	:	confident	=	-0.05 : 1.9
Intimate	:	formal	=	0 : 1.2
Afraid	:	unafraid	=	-1.9 : 1.78

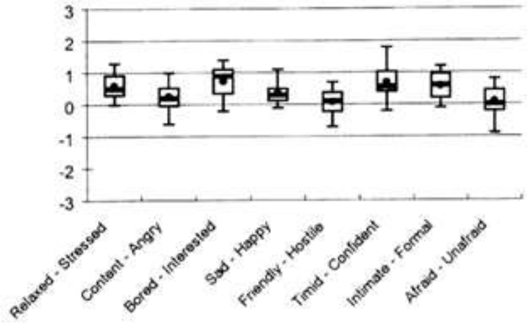

Fig1: parameter ratio test over mood

Thus by the above certain consideration, parameters of voice is taken under analysis and indications can be made. This strategy can mainly used for palliative care in terms medical forms

3. EXPERIMENTAL SETUP

Mood analysis through utterance level is implemented using python 2.7. various parameters of voice like:

- Mel spectrogram
- Harmonic percussive
- Chromagram
- Mel frequency cepstral
- Beat tracking
- Beat-synchronous features aggregation

All these parameters uses some libraries of python like numpy , matplotlib , ipython , librosa.

Numpy : is numerical python for some mathematical operations. It is a fundamental package

Matplotlib : is a plotting library for python programming and used for displaying the output

Ipython : is an interactive command line terminal for python. IPython.display for audio output

Librosa : is an audio and music analysis package in python and is used for audio and the display module for visualization.

3.1 importing the audio file:

The audio file is taken into the process through:

```
audio_path= librosa.util.example_audio_file --1
```

by this form the input audio is taken into the system of process and the parameters are analysed .

3.2 mel spectrogram:

This first step will show how to compute a Mel spectrogram from an audio waveform and will display a mel-scaled power (energy-squared) spectrogram . Convertion to log scale (dB) is made with peak power(max) as reference. sample rate and hop length parameters are used to render the time axis. Mel scale is displayed on spectrogram. Then the plots are titled and figure layout is made compact.

```
S = librosa.feature.melspectrogram

log_S = librosa.power_to_db(S, ref=np.max)

librosa.display.specshow(log_S, sr=sr, x_axis='time', y_axis='mel')

plt.title('mel power spectrogram')

plt.colorbar(format='%+02.0f dB')

plt.tight_layout()
```

3.2 Harmonic-percussive source separation:

Abstract sounds can broadly be classified into two classes. Harmonic sound on the one hand side is what we perceive as pitched sound and what makes us hear melodies and chords. Percussive sound on the other hand is noise-like and usually stems from instruction. The aim of the harmonic/percussive separation is to decompose the original music signal to the harmonic(pitched instrument) and the percussive(non pitched instrument)

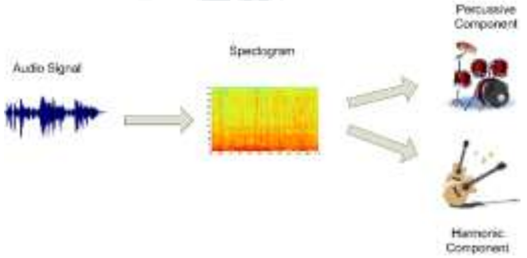

Fig2: Harmonic-percussive source separation

```
y_harmonic, y_percussive = librosa.effects.hpss(y)

S_harmonic=librosa.feature.melspectrogram(y_harmonic, sr=sr)

S_percussive=librosa.feature.melspectrogram(y_percussive, sr=sr)

log_Sh = librosa.power_to_db(S_harmonic, ref=np.max)

log_Sp= librosa.power_to_db(S_percussive, ref=np.max)

plt.figure(figsize=(12,6))

plt.subplot(2,1,1)

librosa.display.specshow(log_Sh, sr=sr, y_axis='mel')

plt.title('mel power spectrogram (Harmonic)')

plt.colorbar(format='%+02.0f dB')
```

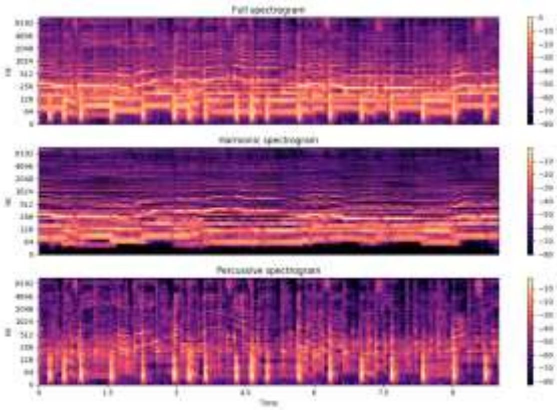

Fig:3 separation of spectrogram

3.3 chromagram:

It is used extract Chroma features to represent pitch class information. chromagram features is powerful representation of music audio in which the entire spectrum is projected onto 12bins representing the 12 distinct chroma of the musical octave. To Display the chromagram: the energy in each chromatic pitch are classified as a function of time

```
C = librosa.feature.chroma_cqt(y=y_harmonic, sr=sr)
plt.figure(figsize=(12,4))

librosa.display.specshow(C, sr=sr, x_axis='time', y_axis='chroma', vmin=0, vmax=1)
plt.title('Chromagram')
plt.colorbar()

plt.tight_layout()
```

3.4 Mel frequency cepstral :

Mel-frequency cepstral (MFCC) is coefficients are commonly used to represent texture or timbre of sound. Next is to extract the top 13 mel-frequency cepstral coefficients. In MFCC , a signal goes through a pre-emphasis filter, then gets sliced into frames and a window function is applied to each frame. Fourier transform is on each frame and power spectrum is calculated and subsequently the filter bank is computed.

```
mfcc        = librosa.feature.mfcc(S=log_S, n_mfcc=13)

delta_mfcc  = librosa.feature.delta(mfcc)

delta2_mfcc = librosa.feature.delta(mfcc, order=2)

plt.subplot(3,1,1)

librosa.display.specshow(mfcc)

plt.ylabel('MFCC')

plt.colorbar()

librosa.display.specshow(delta_mfcc)

plt.ylabel('MFCC-$\Delta$')

plt.colorbar()
```

3.5 Beat tracking:

The beat tracker returns an estimate of the tempo (in beats per minute) and frame indices of beat events. The input can be either an audio time series (as we do below), or an onset strength envelope calculated by

```
librosa.onset.onset_strength()
```

percussive component is used for this part. By default, the beat tracker will trim away any leading or trailing beats that don't appear strong enough. To disable this behavior
:

call beat_track() with trim=False.

```
plt.figure(figsize=(12, 6))

tempo, beats =librosa.beat.beat_track(y=y_percussive, sr=sr)

plt.figure(figsize=(12,4))

librosa.display.specshow(log_S, sr=sr, x_axis='time', y_axis='mel')
```

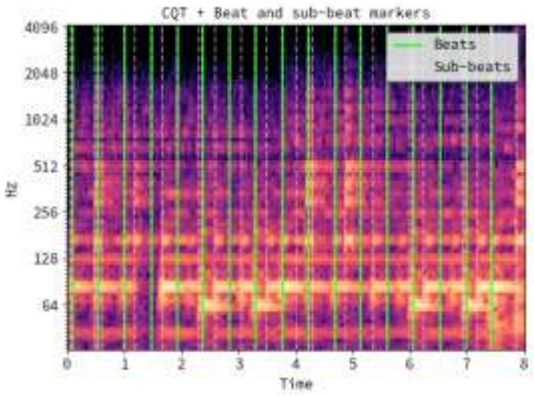

Fig:4 Beat Tracking

3.6 Beat synchronous features aggregation:

In Beat-synchronous feature aggregation, Once we've located the beat events, we can use them to summarize the feature content of each beat. This can be useful for reducing data dimensionality, and removing transient noise from the features. feature.sync will summarize each beat event by the mean feature vector within that beat:

```
M_sync = librosa.util.sync(M, beats)
```

Let's plot the original and beat-synchronous features against each other

```
plt.subplot(2,1,1)

librosa.display.specshow(M)
```

Beat synchronization is flexible. Instead of computing the mean delta-MFCC within each beat, let's go with beat-synchronous chroma. We can replace the mean with any statistical aggregation function, such as min, max, or median.

```
C_sync = librosa.util.sync(C, beats, aggregate=np.median).
```

4. FURTHER PROCESS

All these method can finally be taken under k-nearest neighbor and Deep neural network algorithms for classification. Alert signals can be made through cloud techniques.

4.1 Deep neural network:

Deep neural network (DNN) has more than one hidden layers between its input and output. DNN is a feed-forwarded neural network of AI. High-level representation can be learned from the raw features and data can be effectively classified. It provides proper trained data and correct training strategy. DNN plays a vital role in speech emotion recognition.

4.2 k-nearest neighbors algorithm (k-NN):

k-nearest neighbors algorithm (k-NN) is a parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. k-NN pattern classifier class assignments. This new approach is able to identify 79% of available usable speech segments with 21% false alarms and it requires lesser amount of data to make accurate decisions.

5. USES AND APPLICATIONS

This method of finding the mood or condition of a person through voice is a venturing idea where the usefulness of this idea is high, and will share its uses with many sectors from medical to information technologies. This process could be included with video that is, the face recognition by using its own API's merged with voice and can be developed as a software application in future industry, so that it will depict the condition of the person with whom we are interacting with. It will be more and more useful for person abroad with their relatives and parents in their hometown. Person at far away places can easily understand their bound members condition even if he/she do not expresses, which is the revolutionary usefulness of this process. This can also help the salesperson or the entrepreneur of a particular commodity whether the customer of him is satisfied with its service or not. Thus, this process will provide a lively and most usefulness to the environment. A dumb video can detect only 70% of the mood or condition of the person enacting similarly from the voice of a person 75% of the mood or condition could be detected. When both audio and video is combined 95% of the mood of the person could be detected.

6. CONCLUSION

The study over this field infers us the knowledge, that this techniques is yet play the key role in technical field. Voice and face detection will play a wise role in upcoming equipment and systems. The authentication processes also highly lay their concern over this recognition formula. This idea may gain it's assert over the vast field of computer science and other related branches to lay its root firm to become the indispensable one of the future world. The reason for evolution of all these technique is just for the advancement and reduction to time that assist people

7. REFERENCE

1] sequential k-nearest neighbor pattern recognition for usable speech classification - Jashmin K Shan, Brett Smolenki , Robert E Yantorno , Ananth N Iyer.

2] www.vocabulary.com

3]http://nbviewer.jupyter.org/github/librosa/librosa/blob/master/examples/libROSA%20demo.ipynb.

4] Emotion Detection Analysis Through Tone Of User: A Durvey- Ankita Dutta Chowdhuri , Sachin Bojewar.

5]www.neuralnetwork.com